How do Ontology Mappings Change in the Life Sciences?

Anika Gross, Michael Hartung, Andreas Thor, Erhard Rahm
Department of Computer Science, University of Leipzig
Interdisciplinary Center for Bioinformatics, University of Leipzig
{gross,hartung,thor,rahm}@informatik.uni-leipzig.de

April 13, 2012

Abstract

Mappings between related ontologies are increasingly used to support data integration and analysis tasks. Changes in the ontologies also require the adaptation of ontology mappings. So far the evolution of ontology mappings has received little attention albeit ontologies change continuously especially in the life sciences. We therefore analyze how mappings between popular life science ontologies evolve for different match algorithms. We also evaluate which semantic ontology changes primarily affect the mappings. We further investigate alternatives to predict or estimate the degree of future mapping changes based on previous ontology and mapping transitions.

Keywords: mapping evolution, ontology matching, ontology evolution

1 Introduction

Ontologies have become increasingly important in the life sciences [4, 18]. They are used to semantically annotate molecular-biological objects such as proteins or pathways [27]. Different ontologies of the same domain often contain overlapping and related information. For instance, information about mammalian anatomy can be found in NCI Thesaurus [19] and Adult Mouse Anatomy [1]. Ontology mappings are used to express the semantic relationships between different but related ontologies, e.g., by linking equivalent concepts of two ontologies.

Mappings between related ontologies are useful in many ways, in particular for data integration and enhanced analysis [21, 15]. In particular, such mappings are needed to merge ontologies, e.g., to create an integrated cross-species anatomy ontology such as the Uber ontology [29]. Anatomy ontology mappings may also be useful to transfer knowledge from different experiments between species [3]. Furthermore, mappings can help finding objects with similar ontological properties as interesting targets for a comparative analysis. Ontology curators can further find missing ontology annotations and get recommendations for possible ontology enhancements based on mappings to other ontologies.

Ontologies underly continuous modifications so that new ontology versions are released periodically [13]. New versions typically incorporate enhanced

knowledge, such as additional concepts, relationships, and attribute values. Existing information can also be revised or even deleted. Such ontology changes can invalidate previously determined ontology mappings so that they may have to be re-determined to remain useful. Unfortunately, determining ontology mappings is an expensive process even with the help of semi-automatic ontology matching techniques [7, 24] that still involve a manual verification of correspondences and a parametrization effort. The importance on determining and adapting ontology mappings is underlined by the popular Ontology Alignment Evaluation Initiative (OAEI) [22]. OAEI provides real-world test data sets, in particular for matching the Adult Mouse Anatomy Ontology against the anatomy part of NCI Thesaurus. Unfortunately, the reference mapping of the anatomy task is based on 5 year old ontology versions[1] so that its quality for the current ontology versions remains unclear.

The evolution of ontology mappings has received very little attention so far, especially for the life science domain. For example it is unknown to what degree and how mappings between popular life science ontologies change and how ontology changes affect ontology mappings. There are many ways to compute mappings and it is not clear to what degree different match methods result in differently stable ontology mappings. Finally, we would like to investigate to what degree one can predict future mapping changes based on previously observed ontology and mapping changes. Such information is expected to be useful for deciding about whether a previous ontology mapping is still reliable and up-to-date or whether one has to perform an expensive adaptation of the mapping.

To address these questions and issues we make the following contributions:

- We introduce a generic model for ontology and mapping evolution as well as for their inter-dependencies. The model supports analyzing the impact of ontology evolution on mapping evolution, e.g., what ontology changes lead to the addition or deletion of correspondences in the mapping. (Sec. 3)
- We apply our model to three life science scenarios and evaluate how mappings between popular life science ontologies evolve. We also investigate mapping evolution for different match techniques. (Sec. 4)
- We propose and evaluate two approaches to estimate the number of mapping changes based on previous ontology and mapping changes. (Sec. 5)

In Sec. 2 we present preliminaries and outline the general scenario. We describe related work in Sec. 6 and conclude in Sec. 7.

2 Preliminaries

2.1 Ontology, Mapping, and Matching

In general an **ontology** $O = (C, R, A)$ consists of concepts C which are interrelated by directed relationships R. Each concept has an unambiguous identifier such as an accession number. A concept typically has further attributes $a \in A$ to describe the concept, e.g., name, synonyms, or definition. A relationship $r \in R$

[1] As of 2012, the current reference ontology mapping has been created in 2007.

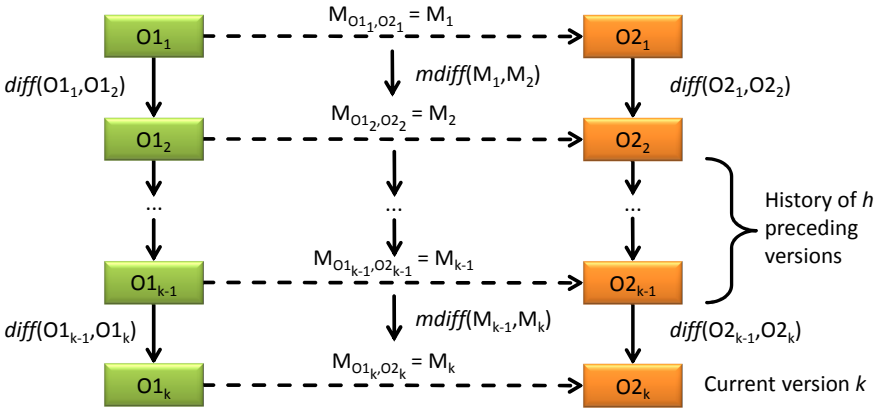


Figure 1: General evolution scheme with multiple ontology and mapping versions

forms a directed connection between two concepts and has a specific type, e.g., is_a or part_of. An **ontology mapping** $M_{O1,O2}$ is a set of correspondences $(c1, c2)$ whereby each correspondence interconnects two concepts $c1 \in O1$ and $c2 \in O2$ of the two ontologies. The mapping semantics depends on the intended use case but we assume that all correspondences of a mapping express the same semantic type, e.g., is-equivalent-to or is-related-to.

Since a purely manual creation of ontology mappings is a tedious and labor-intensive task such mappings are usually determined by semi-automatic **ontology matching** techniques (see Sec. 6 for Related Work). Most matching approaches are metadata-based, i.e., they use the ontology representations themselves to find related concepts, in particular the names of concepts and contextual information like the names of the parent or child concepts within the ontologies. In our evaluation, we will analyze mapping changes for three typical metadata-based matchers (Sec. 4).

2.2 Versioning Scheme

We define an **ontology version** $O_v = (C_v, R_v, A_v)$ as a snapshot of an ontology O released at a specific point in time. For simplicity we enumerate the versions with ascending numbers $v = 1, 2, \ldots$ rather than using the actual release dates.

Ontology changes affect previously determined ontology mappings so that these mappings should be continuously adapted. Fig. 1 illustrates the general versioning scheme we adopt in this paper. There is a series of versions ($v = 1 \ldots k$) for a pair of ontologies $O1$ and $O2$ that are connected by an ontology mapping $M_{O1,O2}$. For simplicity we determine ontology mappings only between ontologies of the same version number, i.e., we create mappings M_v only between ontology versions $O1_v$ and $O2_v$ referring to the same specific point in time.

The difference between two ontology and mapping versions is denoted by $diff(O_v, O_{v+1})$ and $mdiff(M_v, M_{v+1})$, respectively. The next section explains $diff$ and $mdiff$ in more detail.

Change operation	Type
Insertion of a new concept to O_{v+1} Insertion of a subgraph to a concept Insertion of new relationship in O_{v+1} Addition of an attribute (to an existing concept) Mark concept as non-obsolete	Information extension
Deletion of a concept in O_v Removal of a subgraph Deletion of an relationship in O_v Deletion of an existing attribute Mark concept as obsolete	Information reduction
Split concept of O_v into multiple concepts in O_{v+1} Merge concepts of O_v into a single concept in O_{v+1} Concept substitution Move concept Change attribute value	Information revision

Table 1: COntoDiff change operations (including their categorization in three groups) for ontology evolution $O_v \mapsto O_{v+1}$.

3 Change Model for Ontologies and Mappings

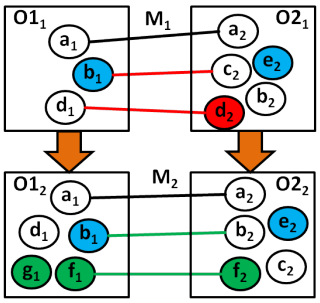
We first describe our change model for ontologies and mappings and categorize the changes into different groups. We also propose simple change ratio indicators to assess the evolution intensity between successive ontology and mapping versions. We then propose indicators to assess the impact of ontology changes on ontology mappings.

3.1 Ontology Changes

We start by defining what changes can occur between successive ontology versions O_v and O_{v+1}. Our model is based on the COntoDiff algorithm described in [12]. COntoDiff computes the difference $diff(O_v, O_{v+1})$ between an old and a new version of an ontology and consists of the set of change operations that – when applied to O_v – transform the old into the new version. Basic change operations are concept and attribute additions or deletions. COntoDiff also determines more complex changes such as merging or splitting of concepts or the addition/deletion of subgraphs.

Table 1 lists all considered change operations and additionally categorizes them into one of three groups. The first group contains information extending operations that add information in O_v such as new concepts, relationships or attribute values. The second group, information reduction, includes change operations that remove information from O_v. All other operations including split and merge changes belong to the revise group.

For a quantitative change analysis we assign concepts both from O_v and O_{v+1} based on their change operations to one of the following sets:



Mapping	Ontologies	EXT	RED	REV
		$\{f_1, g_1\} \cup \{f_2\}$	$\emptyset \cup \{d_2\}$	$\{b_1\} \cup \{e_2\}$
ADD	$\{(b_1,b_2),(f_1,f_2)\}$	2/3	0	1/2
DEL	$\{(b_1,c_2),(d_1,d_2)\}$	0	1	1/2

Figure 2: **left:** Example evolution of two ontologies and a mapping. Concepts b_1 and e_2 have been revised, $d_2 \in O2$ has been removed, and g_1, f_1, and f_2 have been added during the evolution from version $v = 1 \mapsto 2$. The mapping change between $O1$ and $O2$ comprises two new correspondences $((b_1, b_2)$, $(f_1, f_2))$ and two removed correspondences $((b_1, c_1)$, $(d_1, d_2))$. **right:** Impact matrix of ontology and mapping changes.

- **Extension set:** $Ext(O_{v \mapsto v+1})$ = set of concepts in $O_v \cup O_{v+1}$ where all concept-related change operations are information extending.
- **Reduction set:** $Red(O_{v \mapsto v+1})$ = set of concepts in $O_v \cup O_{v+1}$ where all concept-related change operations are information reducing.
- **Revision set:** $Rev(O_{v \mapsto v+1})$ = set of concepts in $O_v \cup O_{v+1}$ that are involved in at least one change operation but belong neither to Ext nor to Red. Each concept is thus related to a revise operation or is related to both extending and reducing operations.

All other concepts remain unchanged, i.e., they are not affected by any change operation. Fig. 2 illustrates an evolution example for two ontologies $O1$ and $O2$. For example, the evolution from $O2_1$ to $O2_2$ might contain three change operations: insertion of concept f_2, deletion of concept d_2, and an attribute value change for concept e_2. The three concepts are thus assigned to Ext, Red, and Rev, respectively, i.e., $Ext(O2_{1 \mapsto 2}) = \{f_2\}$, $Red(O2_{1 \mapsto 2}) = \{d_2\}$, and $Rev(O2_{1 \mapsto 2}) = \{e_2\}$. All other concepts of Fig. 2 are not affected by the change operations.

The size of the three concept sets Ext, Red, and Rev quantitatively characterizes the degree of change during the evolution from O_v to O_{v+1}. We therefore define the **ontology change ratio** as follows:

$$OCR(O_{v \mapsto v+1}) = \frac{|Ext(O_{v \mapsto v+1}) \cup Red(O_{v \mapsto v+1}) \cup Rev(O_{v \mapsto v+1})|}{|O_v \cup O_{v+1}|}$$

The ontology change ratio for $O2$ of our running example (Fig. 2) is thus $OCR(O2_{1 \mapsto 2}) = |\{f_2, d_2, e_2\}|/|\{a_2, b_2, c_2, d_2, e_2, f_2\}| = 0.5$.

3.2 Mapping Changes

For ontology mapping evolution we employ a simple model that distinguishes between the addition and deletion of correspondences. Thus, between two con-

secutive mapping versions M_v and M_{v+1} we consider whether a new correspondence has been added (*Add*) or a previous one has been removed (*Del*). We group changed correspondences into the following sets:

- **Addition set:** $Add(M_{v\mapsto v+1}) = M_{v+1} \backslash M_v$
- **Deletion set:** $Del(M_{v\mapsto v+1}) = M_v \backslash M_{v+1}$

All other correspondences appear in both mapping versions and are thus unchanged. Based on the introduced sets we define the **mapping change ratio** as follows:

$$MCR(M_{v\mapsto v+1}) = \frac{|Add(M_{v\mapsto v+1}) \cup Del(M_{v\mapsto v+1})|}{|M_v \cup M_{v+1}|}$$

In the example of Fig. 2 there are two new correspondences, i.e., $Add(M_{1\mapsto 2}) = \{(b_1, b_2), (f_1, f_2)\}$. and two deleted correspondences, (b_1, c_2) and (d_1, d_2). Since there is one unchanged correspondence (a_1, a_2), the mapping change ratio $MCR(M_{1\mapsto 2})$ equals 4/5.

3.3 Impact of Ontology on Mapping Changes

To determine how ontology changes influence or trigger mapping changes it is useful to interrelate the different kinds of ontology changes and mapping changes. For this purpose, we interrelate the three sets of changed concepts (*Ext*, *Red*, *Rev*) with the two sets of changed correspondences (*Add*, *Del*). We will define six corresponding indicators and use them for both analyzing mapping evolution (see Sec. 4) as well as for predicting mapping changes for new ontology versions (see Sec. 5).

The **impact ratio** is the share of changed concepts that actually had an impact on the correspondences. For any set of ontology changes O_{Ch} (*Ext*, *Red*, or *Rev*) and mapping changes M_{Ch} (*Add* or *Del*) it is defined as follows:

$$IR(O_{Ch}, M_{Ch}) = \frac{|\{c \in O_{Ch} | \exists c' : (c, c') \in M_{Ch} \vee (c', c) \in M_{Ch}\}|}{|O_{Ch}|}$$

For example, to determine which fraction of additive ontology changes led to new correspondences we determine the impact ratio for $O_{Ch} = Ext(O1_{1\mapsto 2}) \cup Ext(O2_{1\mapsto 2})$ and $M_{Ch} = Add(M_{1\mapsto 2})$. For the example in Fig. 2, two (f_1 and f_2) out of the three *Ext*-concepts appear in the set of added correspondences, i.e., the changes in these two concepts had an impact on the mapping. Therefore $IR(Ext, Add)$ equals $\frac{2}{3}$.

One would expect that *Ext* concepts mostly lead to correspondence additions whereas *Red* concepts usually account for correspondence deletions. However, as we will see in our evaluation (see Sec. 4), *Ext* concepts may also trigger correspondence deletions and *Red* concepts may lead to new correspondences depending on the match technique.

4 Analysis of Mapping Evolution

After introducing the experimental setup, we analyze ontology and mapping evolution for different life science scenarios. We then compare mapping evolution for different match strategies and evaluate the impact of ontology changes on mapping changes.

	ontologies		*Name 0.6*		*NameSyn 0.6*		*NameSyn 0.8*		*Context 0.6*	
	$\|C_{2006\text{-}06}\|$	growth	$\|M_{2006\text{-}06}\|$	growth	$\|M_{2006\text{-}06}\|$	growth	$\|M_{2006\text{-}06}\|$	growth	$\|M_{2006\text{-}06}\|$	growth
Anatomy	8,806	1.1	1,496	1.1	1,636	1.1	1,264	1.1	1,272	1.0
Molecular Biology	18,974	1.6	852	1.1	1,531	1.7	251	1.6	465	1.6
Chemistry	69,005	1.7	1,353	3.9	3,242	3.2	1,930	3.7	277	6.1

Figure 3: Ontology and mapping growth factors. Number of concepts ($|C_{2006-06}|$) and number of mapping correspondences ($|M_{2006-06}|$) in the first considered version. $|C|$ is the sum of domain and range ontology size for each match problem. Growth factors compare the first (2006-06) and last (2010-12) considered version.

4.1 Setup

We consider three mapping scenarios:

- *Anatomy*: map Adult Mouse Anatomy Ontology (MA) to the anatomy part of NCI Thesaurus (NCITa)
- *Molecular Biology*: map the two Gene Ontology[9] sub-ontologies Molecular Functions (MF) and Biological Processes (BP)
- *Chemistry*: map Chemical Entities of Biological Interest (ChEBI) [5] to NCI Thesaurus (NCIT)

For each input ontology we map 10 versions on a half year basis between 2006-06 and 2010-12 with each other. We use the following meta-data based matchers to compute the confidence (similarity) for any concept pair of two ontologies:

- *Name*: String (trigram) similarity of concept names
- *NameSyn*: Maximal string (trigram) similarity of names and synonyms
- *Context*: String (trigram) similarity of the concatenated parent, concept, and children names

In this study we focus on the evolution of ontology mappings and do not evaluate the quality of matching. The choice of match strategies is based on previous studies where matching on concept names and synonyms achieved high quality especially for anatomy ontologies [10, 11]. To obtain precise results we need to select the most likely correspondences exceeding a certain confidence threshold. We applied a default confidence threshold of *0.6*; for the *NameSyn* matcher, we also considered a stricter threshold of *0.8*. Moreover, for each input ontology concept, we only select the top correspondences in a small delta range (MaxDelta selection [6]).

4.2 Ontology and Mapping Evolution

Fig. 3 gives an overview about the ontology and mapping sizes as well as their growth between June 2006 and Dec. 2010. For *Anatomy*, the combined size of concepts in domain and range ontology ($|C|$) grew only slightly by a factor 1.1 to almost 10,000 concepts. By contrast, $|C|$ increased by 60 - 70 % to 30,000 and 120,000 concepts for *Molecular Biology* and *Chemistry*. In two of the three

scenarios (*Anatomy* and *Molecular Biology*), the mappings grow similarly strong as the ontologies while the *Chemistry* mappings grew by up to a factor 6. The especially high mapping growth for the *Context* matcher seems influenced by its very small mapping size which in turn is caused by its need to find similar names not only for the concepts but also for their parent and child concepts. Comparing the results for *NameSyn* with two different thresholds, we find that a higher threshold produces smaller mappings and achieves only a relatively small coverage, especially for *Molecular Biology*. For *Molecular Biology*, the *Name* matcher proved to determine the most stable mappings.

Fig. 4(a) shows ontology change factors (see Sec.3.3) between succeeding versions for the three domains during the 5-year observation period. For *Anatomy* there were only few changes over time compared to the other two domains. *Molecular Biology* shows high change rates until 2007 (nearly 40%). From 2008 on, change rates are comparable to those of *Chemistry* (around 20%). Fig. 4(b) illustrates more detailed mapping evolution results for *NameSyn 0.6* in *Molecular Biology*. In general, correspondence additions dominate leading to a final mapping size of more than 2,500 correspondences. But there has also been a considerable number of deletions. In 2007-12 nearly 500 correspondences were removed from the mapping. This shows that there can be very heavy mapping changes.

4.3 Comparison of Match Strategies

To analyze the mapping stability for different match strategies in more detail, we examine a possible correlation between ontology and mapping changes over time. We therefore compute ontology and mapping change factors for all three match scenarios and the four match strategies (Fig. 5 a-c). For *Anatomy*, ontologies and mappings only slightly changed (see y-axis range), while the other two scenarios experience a surprisingly high degree of mapping changes between 10 and 80 %. Except for *Chemistry* we observe a strong correlation between the ontology change factor (black continuous line) and the mapping change factors of the different match strategies(colored dashed lines). The *Name* matcher was relatively stable in general while the *Context* matcher was most heavily influenced by ontology evolution. This especially holds for *Chemistry* where 80% of the *Context* mappings changed in 2008. The reason for the relative instability of *Context* is mainly in its use of more ontological information that can change, i.e., changes on both parent and child concepts have an influence. For instance,

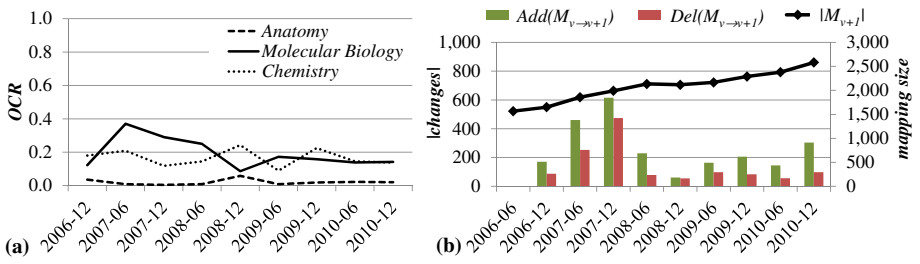


Figure 4: (a) Ontology change factors. (b) Mapping evolution for *NameSyn 0.6* matcher in *Molecular Biology* example.

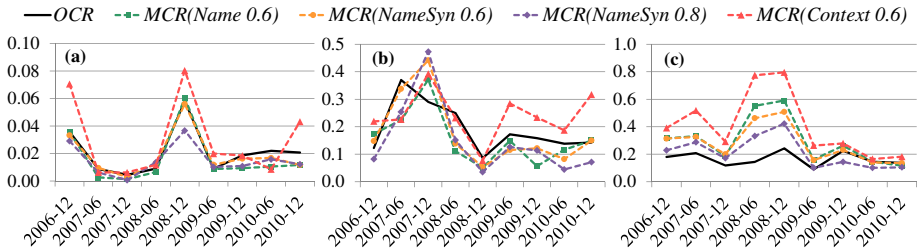


Figure 5: Ontology and mapping change factors for three life science domain examples (a) *Anatomy*, (b) *Molecular Biology*, (c) *Chemistry*

moving a concept from one parent concept to another might completely change a concept's context. For *Molecular Biology* the mappings, (especially *NameSyn*), changed heavily in 2007-12, although the maximum ontology evolution already occurred in 2007-06. This results from successive modification of GO-BP and GO-MF in 2007. The combined changes in both sub-ontologies seem to have led to numerous mapping changes in 2007-12.

4.4 Impact of Ontology on Mapping Changes

Fig. 6 illustrates the real impact of ontology changes (*Ext*, *Red*, *Rev*) on mapping changes (*Add*, *Del*). We exemplarily show results for *NameSyn 0.6* and computed the average over all versions. The table shows the number of changed concepts as well as the ratio having impact on mapping changes (IR). First, we can observe that a high number of ontology extensions, reductions and revisions has no impact on the ontology mappings (>80%). This is due to a limited match coverage since changed ontology parts that are not covered by the ontology mapping do not result in mapping changes. Second, extending ontology changes (*Ext*) primarily cause correspondence additions and no or only few correspondence deletions for all three scenarios. Third, *Red* concepts are primarily involved in correspondence deletions but also in some additions. The latter might result from specific matcher characteristics. Imagine a concept loses a synonym and also the correspondence based on this synonym. This can enable a new correspondence by relating the concept to another one than before. Thus, a synonym deletion can lead to a correspondence deletion and addition in one evolution step. Finally, revised concepts (*Rev*) trigger both, *Add* and *Del*. This

	|Ext|	**IR_{Ext}**		**|Red|**	**IR_{Red}**		**|Rev|**	**IR_{Rev}**	
		→Add	**→Del**		**→Add**	**→Del**		**→Add**	**→Del**
Anatomy	**95**	18.7%	0.1%	**7**	0.0%	7.8%	**89**	6.8%	4.1%
Molecular Biology	**2,359**	4.6%	0.7%	**223**	2.4%	8.8%	**2,209**	3.5%	2.1%
Chemistry	**8,377**	11.7%	1.2%	**366**	3.5%	5.3%	**6,441**	8.1%	4.0%

Figure 6: Impact of ontology concept changes (*Ext*, *Red*, *Rev*) on mapping changes (*Add*, *Del*) for *NameSyn 0.6*. Average values for absolute change number ($|Ext|$, $|Red|$, $|Rev|$) and impact association ratios ($IR(O_{Ch}, M_{Ch})$ displayed as percentage) over all considered versions

i→i+1	*|Ext|*	*IR(Ext,Add)*	*|Red|*	*IR(Red,Add)*	*|Rev|*	*IR(Rev,Add)*	*|Add|*
1→2	60	0.3	10	0.1	15	0.2	20
2→3	30	0.4	2	0	10	0.1	10
3→4	**40**	**?**	**4**	**?**	**12**	**?**	**?**

Table 2: Example prediction scenario.

is intuitive since revised concepts might have been extended and reduced in one evolution step (e.g., attribute addition and deletion). In general, ontology revisions account for a high share of mapping changes while deletions play only a minor role.

4.5 Summary

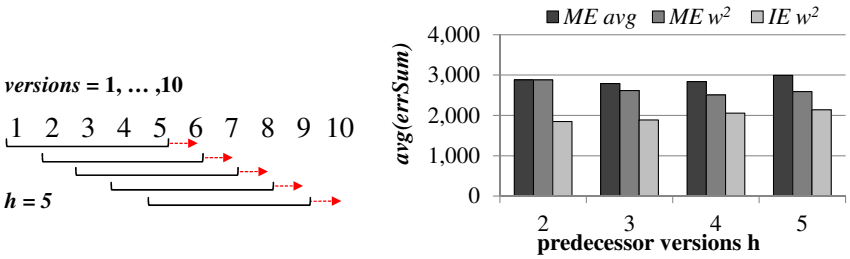
We evaluated ontology and mapping evolution for three real-world life science domains (*Anatomy*, *Molecular Biology* and *Chemistry*) and took four match-strategies into account. The analysis results show that especially *Molecular Biology* and *Chemistry* underlie heavy ontology extensions and revisions whereas *Anatomy* is relatively stable. Since existing knowledge is mainly extended or revised, we find only few ontology reducing changes for all domains. Ontology evolution heavily influenced mappings computed by different metadata-based match strategies. Especially, the structural matcher *Context* produced rather unstable results whereas mappings based on the *Name* matcher are relatively stable. As expected, ontology extensions primarily lead to correspondence additions and information reducing ontology changes primarily lead to the removal of correspondences. Ontology revisions play an important role and result in both the addition and deletion of correspondences.

5 Mapping Change Estimation

We now present two methods to estimate the number of changes in a new mapping version. By predicting future mapping changes we can give recommendations to users if it might be necessary to recompute their mappings. This seems especially useful when one must decide about performing an expensive manual mapping adaption or not. We first describe the methods and then comparatively evaluate their quality on our mapping problems.

5.1 Prediction Methods

The general task of estimating mapping changes is the following. After the release of two new ontology versions $O1_k/O2_k$ we like to predict the number of mapping changes ($|Add(M_{k\text{-}1\mapsto k}|$,$|Del(M_{k\text{-}1\mapsto k}|$) which will occur between the mapping versions $M_{k\text{-}1}$ and M_k, i.e., we like to know how strong mapping M is likely to change due to modifications in $O1/O2$. For this estimation we can access the content of the previous h ontology/mapping versions (*v=k-h,. . . ,k-1*) and their diff results. In the following we describe two prediction methods, namely *Mapping-based Estimation (ME)* and *Impact-based Estimation (IE)*. The synthetic example in Table 2 will be used for illustration.



(a) (b)

Figure 7: Prediction analysis (a) Example for predicting the successor version (red dotted line) on the basis of a window of 5 predecessor versions (h=5), (b) Average error sum ($avg(errSum)$) of false predictions for $h = 2 \ldots 5$ for three methods $MEavg$, MEw^2, IEw^2

Mapping-based Estimation In this approach the prediction only uses information about previous mapping changes but not about the underlying ontology changes. The estimation for $|Add(M_{k\text{-}1 \mapsto k})|$ and $|Del(M_{k\text{-}1 \mapsto k})|$ is the weighted average of the number of changes observed in the last h-1 version changes of the mapping. We can use different functions w to weight the version changes:

$$|Add(M_{k\text{-}1 \mapsto k})| = \sum_{v=k-h+1}^{k-1} w_i \cdot |Add(M_{v\text{-}1 \mapsto v})|$$
$$|Del(M_{k\text{-}1 \mapsto k})| = \sum_{v=k-h+1}^{k-1} w_i \cdot |Del(M_{v\text{-}1 \mapsto v})|$$

For our example in Table 2 we like to make a prediction for the number of added correspondences between version 3 and 4 ($|Add(M_{3 \mapsto 4})|$) using the versions 1–3 (h=3). We use a quadratic weighting function with the following weights for the two previous version changes: $\frac{1}{5}$ and $\frac{4}{5}$. We would thus estimate $|Add(M_{3 \mapsto 4})|$ $= \frac{1}{5} \cdot 20 + \frac{4}{5} \cdot 10 = 12$ with the ME method.

Impact-based Estimation The idea behind impact-based estimation is to use knowledge about the impact of ontology on mapping changes to estimate the number of correspondence changes. We assume that the number of added/deleted correspondences can be expressed as a linear combination of the observed ontology changes having an impact:

$$\begin{aligned} |Add(M_{k\text{-}1 \mapsto k})| &= \beta \cdot (\mathsf{agg}(IR(Ext, Add)) \cdot |Ext(O_{k\text{-}1 \mapsto k})| \\ &+ \mathsf{agg}(IR(Red, Add)) \cdot |Red(O_{k\text{-}1 \mapsto k})| \\ &+ \mathsf{agg}(IR(Rev, Add)) \cdot |Rev(O_{k\text{-}1 \mapsto k})|) \end{aligned}$$

$$\begin{aligned} |Del(M_{k\text{-}1 \mapsto k})| &= \beta \cdot (\mathsf{agg}(IR(Ext, Del)) \cdot |Ext(O_{k\text{-}1 \mapsto k})| \\ &+ \mathsf{agg}(IR(Red, Del)) \cdot |Red(O_{k\text{-}1 \mapsto k})| \\ &+ \mathsf{agg}(IR(Rev, Del)) \cdot |Rev(O_{k\text{-}1 \mapsto k})|) \end{aligned}$$

For both formulas we need two specify two parameters. First, we need to determine the impact ratios (IR) which indicate how strong ontology changes will influence the mapping in the current version change. Since we consider the last h-1 version changes, we need to aggregate the observed impact ratios into a common value (agg function), e.g., by a normal or weighted average. Second,

we need to determine the β parameter which performs an error correction on the result. In particular, for each version change we calculate the estimated value using the linear combination formula with the impact ratios observed. We then compare the estimation with the correct result and compute an error ratio between both. We finally take the average of all computed error ratios as our β.

For our example we need to determine three impact ratios. We will use the same quadratic weighting as for *ME* to compute a weighted average. Thus, for $IR(Ext, Add)$ we would determine a value of $\frac{1}{5} \cdot 0.3 + \frac{4}{5} \cdot 0.4 = 0.38$ ($IR(Red, Add) = 0.02$ and $IR(Rev, Add) = 0.12$). We further calculate the error ratio for each version change, e.g., for $1 \mapsto 2$ the estimated result is $0.3 \cdot 60 + 0.1 \cdot 10 + 0.2 \cdot 15 = 22$. A comparison with the correct number of correspondence additions ($|Add(M_{1 \mapsto 2})| = 20$) results in a ratio of $\frac{20}{22} \approx 0.91$ ($2 \mapsto 3 : 0.77$). Thus, the average error ratio over all version changes is $\beta = 0.84$ resulting in an estimation of $|Add(M_{3 \mapsto 4})| = 0.84 \cdot (0.38 \cdot 40 + 0.02 \cdot 4 + 0.12 \cdot 12) \approx 14$.

5.2 Evaluation

We now apply our two estimation methods to predict how many correspondence additions and deletions might occur in a future mapping. We use the same datasets as before (see Sec. 4). For the map-based method (*ME*), we applied two different weight functions: average (*avg*) and quadratic weighting average (w^2). For the Impact-based method (*IE*), we only show results for w^2 since this showed to be more effective. To get an overview how accurate both methods are and how many versions are required for good estimation, we performed the following experiment.

We predict the last five mapping versions using several numbers of predecessor versions ($h = 2 \ldots 5$). Fig. 7(a) exemplarily shows the experimental scenario for $h = 5$. We produce five results per h for the *ME avg*, *ME* w^2 and *IE* w^2 prediction methods. For each h and method we compute an error sum (*errSum*: sum of absolute differences between correct (*CR*) and predicted result (*PR*)) over all prediction results for three matchers (*Name 0.6*, *NameSyn 0.6*, *Context 0.6*) and all match scenarios. To better compare the methods and to study the influence of h we compute average error sums which are displayed in Fig. 7(b). For $h = 2$, *ME avg* and *ME* w^2 produce the same results since they only consider one mapping version diff. For a higher number of predecessor versions ($h > 2$) *ME* w^2 produces smaller errors. Overall *IE* w^2 is more effective than both *ME* methods, i.e., using information about ontology evolution as well, seems to be more informative and thus leads to more accurate results. Especially, only considering the recent past (small h) suffices to make a good estimation with our impact-based method *IE*.

To get an impression how many change operations we predict for *ME* w^2 and *IE* w^2, we selected the following case. Considering the change factors in Fig.5 we would expect that it is hard to predict version 2009-06 based on 2008-06 and 2008-12 for all three match scenarios. In particular, for *Anatomy* and *Chemistry* we see a strong decrease in their change factors whereas for *Molecular Biology* an increase occurred. Fig. 8 shows detailed results of the prediction case. The error rate *err* gives the absolute difference of *PR* and *CR* divided by the respective mapping size for the predicted version ($|M_{2009-06}|$). To get an better overview we illustrate *err* on a red green scale. Overall both methods produce

	matcher	$\|M_{2009\text{-}06}\|$	*AddCorr*					*DelCorr*				
				MEw²		*IEw²*			*MEw²*		*IEw²*	
			CR	*PR*	*err*	*PR*	*err*	*CR*	*PR*	*err*	*PR*	*err*
Anatomy	*Name 0.6*	1,592	8	81	0.05	13	0.00	6	16	0.01	2	0.00
	NameSyn 0.6	1,757	13	85	0.04	14	0.00	8	14	0.00	2	0.00
	Context 0.6	1,285	11	67	0.04	11	0.00	15	40	0.02	6	0.01
Molecular Biology	*Name 0.6*	781	74	22	0.07	54	0.03	54	27	0.03	26	0.04
	NameSyn 0.6	2,166	168	63	0.05	150	0.01	117	81	0.02	66	0.02
	Context 0.6	576	114	21	0.16	50	0.11	72	22	0.09	53	0.03
Chemistry	*Name 0.6*	3,934	555	1,785	0.31	685	0.03	141	724	0.15	269	0.03
	NameSyn 0.6	7,868	1,036	3,010	0.25	1,151	0.01	243	1,193	0.12	442	0.03
	Context 0.6	1,230	237	758	0.42	293	0.05	124	642	0.42	236	0.09

Figure 8: Number of correct and estimated *AddCorr* and *DelCorr* operations using mapping versions $M_{2008-06}$-$M_{2008-12}$ to predict changes in $M_{2009-06}$. Comparison of two methods (MEw^2, IEw^2), for three life science domains and the three matchers. CR (PR) - number of correct (predicted) result, err - error rate on a red (high err) green (small err) scale

relatively good results (green err values) for correspondence deletions. By contrast, estimating additions seems more complicated. *IE* w^2 produces only small errors for additions whereas *ME* w^2 either estimates too high (for *Anatomy* and *Chemistry*) or too low (for *Molecular Biology*) values (yellow to red err values). This is triggered by the previous trend of mapping evolution, as we have seen in Fig.5. Thus, if the pattern of mapping evolution suddenly changes, methods making an estimation solely on the basis of previous mapping changes fail.

By contrast, IEw^2 involves knowledge about ontology evolution as well as its impact on mapping evolution which leads to more accurate prediction results. Especially considering the overall mapping sizes, the predicted results (PR) for *Anatomy* are very close to the correct results (CR) (e.g., 8–13, 13–14, 11–11 for correspondences additions). In general, it seems very difficult to predict mapping changes for *Chemistry* and the *context 0.6* matcher. For *Chemistry* one and the same ontology change factor can lead to mapping changes of different magnitude so that change prediction becomes a complex task. For *context 0.6*, there are several different influences as the evolution of the concept itself, its parents and its children, making it difficult to correctly predict mapping changes.

In general we can recommend that the OAEI *Anatomy* mapping is still feasible and reliable as there were relatively few ontology changes since 2007. Thus, we would expect only few mapping adaptations. By contrast, knowledge in the *Molecular Biology* or *Chemistry* domains changed dramatically in the last 5 years. Thus, mapping adaptation is strongly recommended to obtain useful mappings.

6 Related Work

In the last decade, ontology matching to semi-automatically create ontology mappings has become an active research field (see [7, 23] for overviews). In the life sciences especially the matching of anatomy ontologies [30] and molecular biological ontologies [2] has attracted considerable interest. Most match approaches focus on improving the quality of computed mappings by applying

different matchers (e.g., based on the name/synonyms of concepts, the ontology structure or associated instances) in a workflow-like manner. For comparing available match systems w.r.t. their quality the OAEI [22] provides gold standard mappings, e.g., between MA and NCIT.

Previous work on ontology evolution (see [8, 14] for surveys) focused on ontology versioning [17], the evolution process itself [25] as well as the detection of changes between ontology versions [20]. Few approaches investigate how changes in ontologies should be propagated to dependent artifacts such as instances or annotations. For example, the ontology evolution process proposed in [26] includes a change propagation phase where performed changes are propagated to other ontologies that are based on the modified ontology.

The evolution of ontology mappings has received only little attention so far. In our previous work [13] we studied the evolution of ontologies, annotations and ontology mappings. We analyzed mapping evolution for one match problem and noticed dramatic increases in the number of correspondences especially for instance-based matchers. In a further study [28] we focused on the stability of correspondences created by an instance-based matcher and proposed measures which allow for a classification of (un)stable correspondences.

In contrast to previous work this study focuses on the impact of ontology on mapping changes, i.e., we investigate (1) how ontology mappings change and (2) study how ontology changes correlate with mapping changes for different matchers. Furthermore, we use the knowledge from the correlation between ontology and mapping changes to estimate the cardinality of future mapping changes. The mapping versions under investigation were created with previously evaluated matchers such as name or name/synonym using the GOMMA system [16].

7 Conclusion and Future Work

We studied the evolution of ontology mappings and analyzed the ontology changes triggering mapping changes as well as the influence of different match techniques. Our analysis covered three life science mappings and three match strategies. Furthermore we proposed two prediction methods for estimating the cardinality of future mapping changes. Except for anatomy ontologies, we observed that ontology mappings based on common match strategies using name and synonym information often experience heavy changes. Our prediction methods were quite effective and could reasonably estimate the number of correspondence additions and removals in a new mapping version. In future work, we plan to investigate how known ontology changes can be used to semi-automatically adapt ontology mappings without a completely new mapping determination.

Acknowledgment

This work is supported by the German Research Foundation (DFG), grant RA 497/18-1 ("Evolution of Ontologies and Mappings").

References

[1] Adult Mouse Anatomy. `http://www.informatics.jax.org/searches/AMA_form`.

[2] O. Bodenreider and A. Burgun. Linking the gene ontology to other biological ontologies. In *Proc. ISMB2005 SIG meeting on Bio-ontologies*, pages 17–18, 2005.

[3] O. Bodenreider, T.F. Hayamizu, M. Ringwald, et al. Of mice and men: Aligning mouse and human anatomies. In *Proc. of AMIA Annual Symposium*, 2005.

[4] O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Briefings in bioinformatics*, 7(3):256–274, 2006.

[5] P. De Matos, R. Alcántara, A. Dekker, et al. Chemical entities of biological interest: an update. *Nucleic acids res.*, 38(suppl 1):D249–D254, 2010.

[6] Hong-Hai Do and Erhard Rahm. Coma: a system for flexible combination of schema matching approaches. In *Proceedings of VLDB*, pages 610–621, 2002.

[7] J Euzenat and P Shvaiko. *Ontology matching*. Springer-Verlag New York, 2007.

[8] G. Flouris, D. Manakanatas, H. Kondylakis, et al. Ontology change: Classification and survey. *The Knowledge Engineering Review*, 23(2):117–152, 2008.

[9] Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Res.*, 36(Database Issue):D440–D444, 2008.

[10] A. Ghazvinian, N.F. Noy, and M.A. Musen. Creating mappings for ontologies in biomedicine: Simple methods work. In *Proc. of AMIA Annual Symposium*, 2009.

[11] A. Gross, M. Hartung, T. Kirsten, and E. Rahm. Mapping composition for matching large life science ontologies. In *2nd Intl. Conf. on Biomed. Ontology (ICBO)*, 2011.

[12] M. Hartung, A. Gross, and E. Rahm. Rule-based Generation of Diff Evolution Mappings between Ontology Versions. *CoRR*, abs/1010.0122, 2010.

[13] M. Hartung, T. Kirsten, and E. Rahm. Analyzing the evolution of life science ontologies and mappings. In *Data Integration in the Life Sciences*, pages 11–27, 2008.

[14] Michael Hartung, James F. Terwilliger, and Erhard Rahm. Recent advances in schema and ontology evolution. In *Schema Matching and Mapping*, pages 149–190. Springer, 2011.

[15] V. Jakoniene and P. Lambrix. Ontology-based integration for bioinformatics. In *VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems-ODBIS 2005*, pages 55–58, 2005.

[16] T. Kirsten, A. Gross, M. Hartung, and E. Rahm. Gomma: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. *Journal of Biomedical Semantics*, 2:6, 2011.

[17] M Klein, D Fensel, A Kiryakov, and D Ognyanov. Ontology versioning and change detection on the web. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 247–259, 2002.

[18] P Lambrix, H Tan, V Jakoniene, and L Strömbäck. Biological ontologies. In *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, pages 85–99. Springer Verlag, 2007.

[19] NCI Thesaurus. `http://ncit.nci.nih.gov/`.

[20] N F Noy and M A Musen. Promptdiff: A fixed-point algorithm for comparing ontology versions. In *Proc. of Nat. Conf. on Artificial Intelligence*, pages 744–750, 2002.

[21] N.F. Noy, N.H. Shah, P.L. Whetzel, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids res.*, 37(suppl 2):W170–W173, 2009.

[22] Ontology Alignment Evaluation Initiative. `http://oaei.ontologymatching.org/`.

[23] E. Rahm. Towards Large Scale Schema and Ontology Matching. In *Schema Matching and Mapping*, chapter 1, pages 3–27. Springer, 2011.

[24] E Rahm and P A Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[25] L Stojanovic. *Methods and tools for ontology evolution.* PhD thesis, University of Karlsruhe, 2004.

[26] L Stojanovic, A Maedche, B Motik, and N Stojanovic. User-driven ontology evolution management. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 133–140, 2002.

[27] P.D. Thomas, H. Mi, and S. Lewis. Ontology annotation: mapping genomic regions to biological function. *Current opinion in chemical biology*, 11(1):4–11, 2007.

[28] Andreas Thor, Michael Hartung, Anika Gross, Toralf Kirsten, and Erhard Rahm. An evolution-based approach for assessing ontology mappings - a case study in the life sciences. In *BTW*, pages 277–286, 2009.

[29] UBERON. `http://obofoundry.org/wiki/index.php/UBERON:Main_Page`.

[30] S. Zhang and O. Bodenreider. Experience in aligning anatomical ontologies. *International journal on Semantic Web and information systems*, 3(2):1–26, 2007.