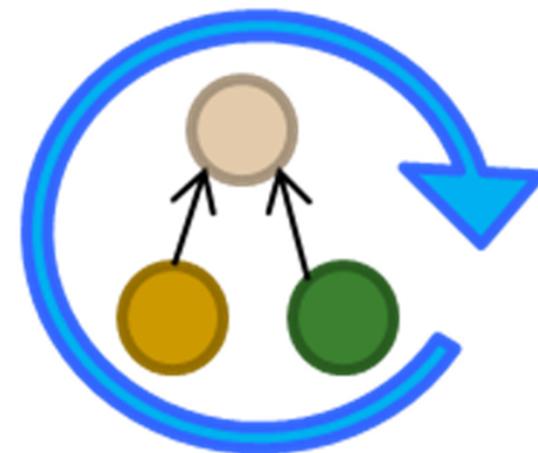

Ontologie-Management

Kapitel 6: Dynamik in Ontologien

Dr. Michael Hartung

Wintersemester 2012/13

Universität Leipzig
Institut für Informatik
<http://dbs.uni-leipzig.de>



Inhalt

- **Einführung**
 - Warum Dynamik?
 - Beispiele und Probleme

- **Aspekte von Dynamik in Ontologien**
 - Prozess-orientierte Adaptierung von Ontologien
 - Berechnung von Diffs zwischen Ontologieversionen
 - Versionierung von Ontologien



Warum Dynamik?

- **Wissen in Ontologien ist nicht statisch**
 - Ständige Änderungen / Anpassungen nötig
 - *Ziel:* möglichst aktueller / korrekter Wissenstand



- **Gründe**
 - Integration von neuem / geändertem Domänenwissen
 - Behebung initialer Designfehler
 - Veränderte Anforderungen seitens der Nutzer
 - Umsetzung neuer Richtlinien
 - Migration zu anderer Ontologiesprache



Probleme durch Dynamik

- **Ausgangspunkt:** Veröffentlichung einer neuen Version
- **Auswirkungen**
 - Ontologie-basierte Daten wie Annotationen / Ontologie-Mappings
 - Sind Annotationen noch gültig?
 - Müssen Ontologie-Mappings angepasst werden?
 - Ontologie-basierte Analysen
 - Stimmen meine Analyseergebnisse noch oder müssen sie revidiert werden?
- **Probleme**
 - Endnutzer oftmals mit Anpassungen konfrontiert
 - Oft keine Unterstützung seitens der Ontologieprovider
 - Größe und Komplexität der Ontologien



Beispiel Szenario

- **Forscher möchte Ontologie in seinen Projekten für Analysen etc. einsetzen**

Ist die Ontologie abgeschlossen oder befindet sie sich noch in Entwicklung?

Was sind die häufigsten Änderungen?

Wie kann ich Änderungen zwischen meiner und einer neuen Version bestimmen?



Wie haben sich einzelne Konzepte im Detail verändert?

Welche Ontologieteile sind stabil oder welche werden häufig angepasst?

Gibt es aufstrebende neue Gebiete innerhalb der Ontologie? Kann ich irgendwo partizipieren?

**Algorithmen, Werkzeuge zur Unterstützung des Forschers:
Änderungsbestimmung + Evolutionsanalyse**

Aspekte von Dynamik in Ontologien

- **Prozess-orientierte Anpassung von Ontologien**
 - Nutzer passt Ontologie nach seinen Vorstellungen / Anforderungen an
 - Sicherung von Konsistenz und Adaptierung abhängiger Daten
- **Berechnung eines Diff zwischen Ontologieversionen**
 - Zentrale immer wiederkehrende Aufgabe
 - Was hat sich zwischen zwei Ontologieversionen verändert?
- **Versionierung großer Ontologien**
 - Gewährleistung eines Zugriffs auf mehrere Ontologieversionen
 - Effiziente Versionierung bei großen Ontologien notwendig



Ontologie-Evolution

Definition

Ontology evolution is the timely adaptation of an ontology to changed business requirements, to trends in ontology instances and patterns of usage of the ontology-based applications, as well as the consistent management/propagation of these changes to dependent elements. (*Stojanovic et al., 2002*)

- Adaptierung (Anpassung) einer Ontologie aufgrund
 - geänderter Anforderungen
 - Trends in den zugehörigen Instanzen
 - Veränderungen im Nutzungsverhalten
- Konsistente Verwaltung / Propagierung durchgeführter Änderungen in abhängige Strukturen
- Ontologie in einen konsistenten Zustand überführen



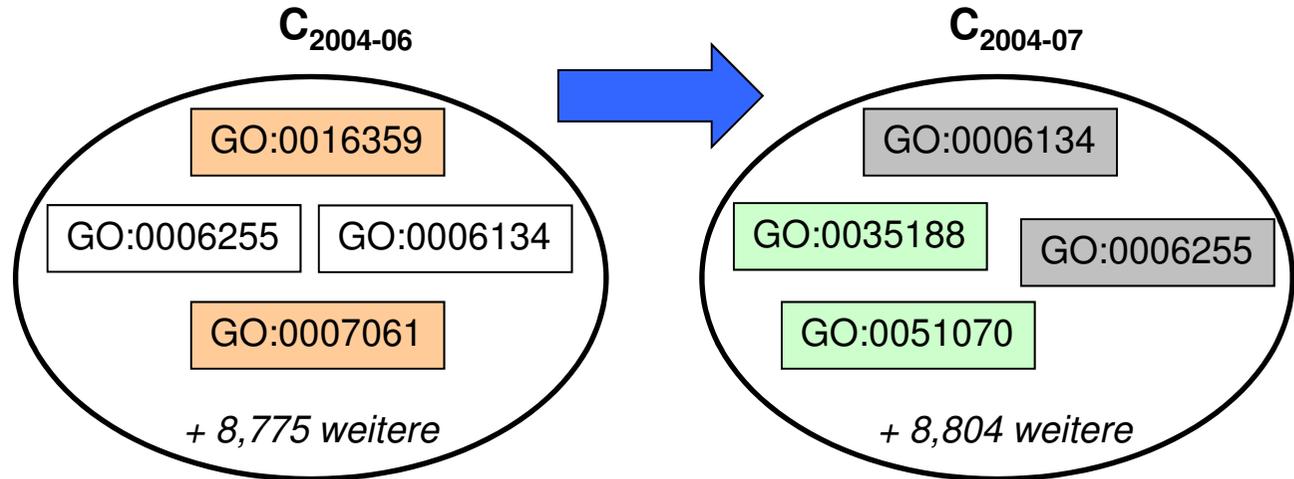
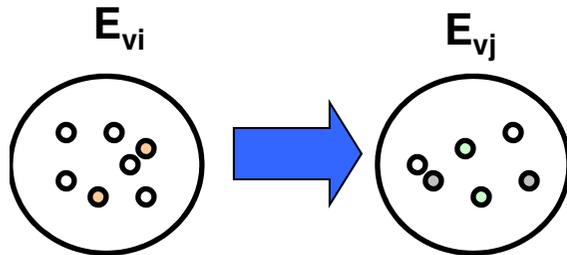
Gibt es überhaupt Evolution?

- **Framework für Evolutionsanalysen**
 - Gemeinsames Evolutionsmodell für Ontologien sowie Mappings
 - Basisänderungen *add / del / toObs* für Elemente (Konzepte, Relationen, ...)
 - Metriken zur Beurteilung der Evolution
- **Umfangreiche quantitative Analysen zwischen 2004 und 2008**
 - 16 Ontologien in den Lebenswissenschaften

Hartung, M., Kirsten, T., Rahm, E.: Analyzing the Evolution of Life Science Ontologies and Mappings. In Proc. Data Integration in the Life Sciences (DILS), 2008



Änderungserkennung: GO-BP 2004-06 → 2004-07



$$add_{vi,vj} = E_{vj} \setminus E_{vi}$$

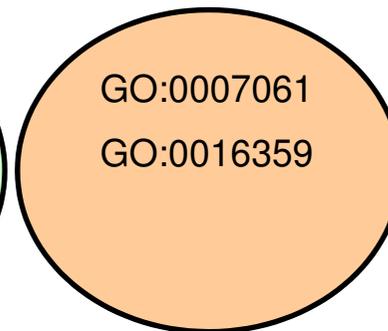
$$del_{vi,vj} = E_{vi} \setminus E_{vj}$$

$$toObs_{vi,vj} = E_{vj,obs} \cap E_{vi,nonObs}$$

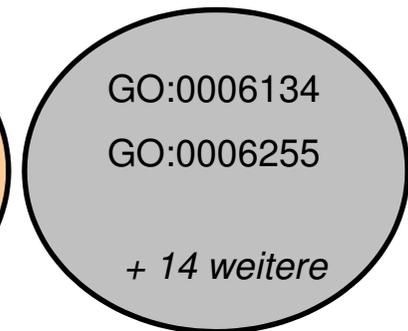
Betroffene Konzepte:



add_{2004-06,2004-07}



del_{2004-06,2004-07}



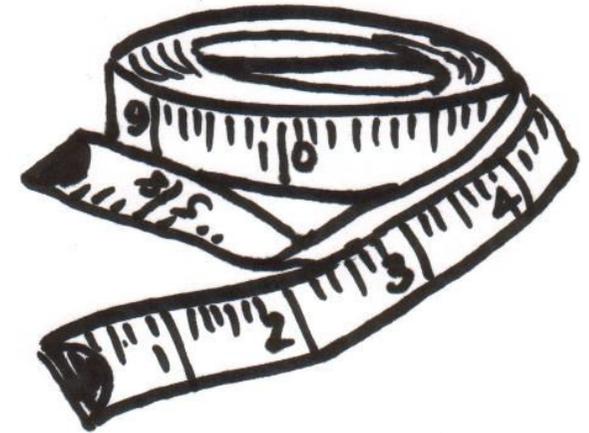
toObs_{2004-06,2004-07}



Framework: Metriken

■ Deskriptive Statistiken

- Anzahl von Elementen in einer Version v_i : $|E_{v_i}|$
- Anzahl innerer Konzepte und Blätter
- Anzahl veralteter und aktiver Konzepte
- Anzahl von is_a, part_of und anderen Relationen
- Anzahl von Pfaden und Pfaden pro Konzept, Pfadlängen



■ Evolutions- und Wachstumsstatistiken

- Wachstum: $growth_{E, v_i, v_j} = |E_{v_j}| / |E_{v_i}|$
- Anzahl geänderter Elemente: Add_{v_i, v_j} , Del_{v_i, v_j} , $ToObs_{v_i, v_j}$
- Anzahl von Änderungen innerhalb einer Zeitperiode p für ein Zeitintervall t : $Add_{p, t}$, $Del_{p, t}$, $ToObs_{p, t}$



Frameworkanwendung

- **Vergleichende Evolutionsanalyse von 16 Ontologien**
 - Ontologien aus verschiedenen Domänen der Lebenswissenschaften
 - Gene Ontology: Funktionen, Prozesse, Komponenten
 - NCI Thesaurus: krebsbezogene Themen
 - Anatomie: Fly, Mouse, Cell, PlantStructure, Zebrafish
 - Andere: ChemicalEntities (ChEBI), MammalianPhenotype, ...
 - 386 Versionen zwischen 2004-05 und 2008-02
 - Größte Ontologien: NCI Thesaurus, GO, ChEBI (>18.000 Konzepte)



Ergebnisse und Beobachtungen *

- **Signifikantes Wachstum und Änderungen in allen Ontologien**
 - Durchschnittliches Wachstum von **1.6**
 - *add* als dominierende Änderung
 - NCI Thesaurus: ~36.000 → ~64.000 Konzepte (627 Einfügungen / Monat)
 - Gene Ontology: ~17.000 → ~26.000 Konzepte (200 Einfügungen / Monat)
 - Auch viele *del* und *toObs* Änderungen
 - ChEBI (62 Löschungen / Monat)
 - Teilweise hohe Instabilitäten in den Ontologien
 - sich entwickelnde Forschungsgebiete

* Detailergebnisse: http://dbs.uni-leipzig.de/lis_ontology_evolution



Ergebnisse und Beobachtungen (2)

■ Wachstum an strukturiertem Wissen

□ *Relationen*

- Anteil von *is_a* ↓ während *part_of* ↑ und *sonstige* ↑, *is_a* dominierend (86% / 7% / 7%)

□ *Innere Konzepte/Pfade*

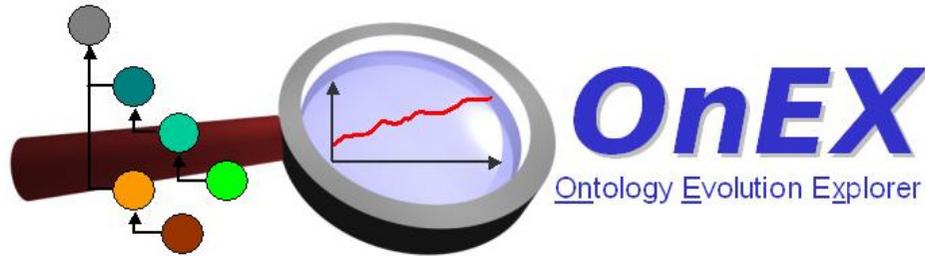
- Anteil innerer Konzepte ↑ → steigende Komplexität
- starkes Wachstum in der Anzahl von Pfaden mit teils sprunghaften Änderungen (Restrukturierung von Ontologien)

■ Aktuell stark verändernd

- Chemical Entities of Biological Interest (ChEBI)
- MammalianPhenotype



OnEX (Ontology Evolution Explorer)



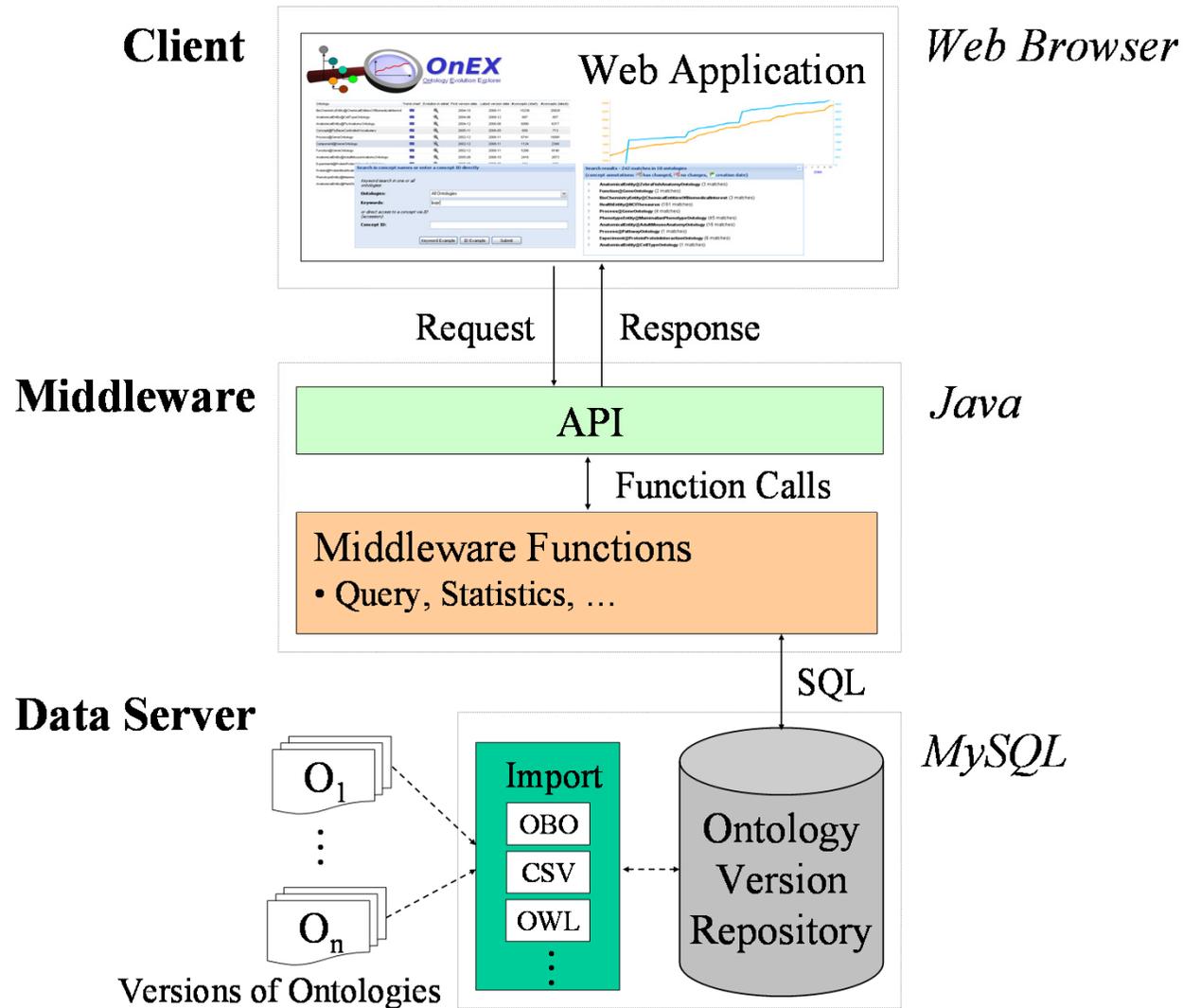
<http://www.izbi.de/onex>

- **Webapplikation für Online-Evolutionsanalysen**
 - Drei Workflows
 - Quantitative Evolutionsanalyse (siehe vorherige Folien)
 - Konzept-basierte Analyse
 - Annotation-Migration
- **3-Schichtenarchitektur**
- **Aktuell: 16 Ontologien mit 850 Versionen**

Hartung, M., Kirsten, T., Groß, A., Rahm, E.: OnEX – Exploring changes in life science ontologies. BMC Bioinformatics 10:250, 2009



3-Schichtenarchitektur



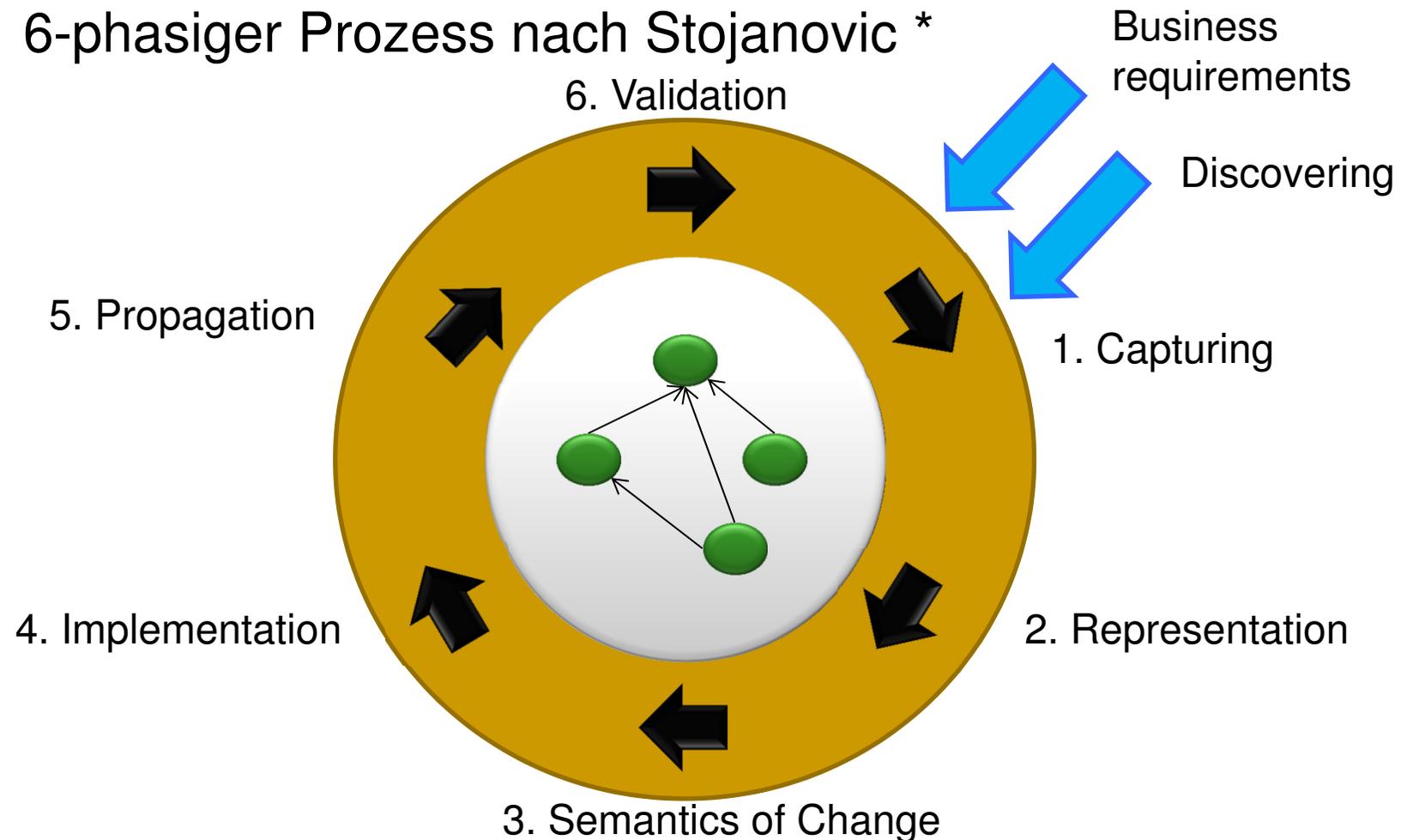
Adaptierung von Ontologien

- *Ausgangspunkt*
 - Konsistente Ontologieversion O_v welche anzupassen ist
 - Menge umzusetzender Anforderungen
- *Ziel*
 - Neue, konsistente Ontologieversion O_v' welche die gesetzten Anforderungen erfüllt
- *Was sollte ein Ontologie-Evolutionsprozess leisten?*
 - Umsetzung/Auflösung von Änderungen sowie die Sicherung der Konsistenz der Ontologie sowie abhängiger Elemente (z.B. andere Ontologien, Instanzen, Mappings)
 - Beaufsichtigung (Eingriff, Kontrolle) durch Nutzer
 - Hinweise für weitere Veränderungen / Verbesserungen



Prozess-orientierte Ontologie-Evolution

- 6-phasiger Prozess nach Stojanovic *



Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-driven Ontology Evolution Management. European Conf. On Knowledge Eng. and Management (EKAW), 2002



1. Capturing

- *Aufgabe*
 - Erfassung der umzusetzenden Änderungen durch Ontologiedesigner
 - aus Anforderungen direkt (business requirements)
 - durch automatische Vorschläge des Systems (discovery)

- *Beispiel*
 - Löschen eines Konzepts C



2. Change Representation

- *Aufgabe*
 - Umsetzung (Transformation) der erfassten Änderungen aus Phase 1 in eine formale, maschinenverständliche Form
 - Unterscheidung zwischen *einfachen* (elementaren) und *komplexen* (composite) Änderungen
 - *Einfach*: Hinzufügen, Löschen von Konzepten, Properties, Axiomen und Beziehungen
 - *Komplex*: Mergen oder Verschieben von Konzepten, Extraktion oder Kopieren von Konzepten, ...

- *Beispiel*
 - *Delete_Concept(C)*



3. Semantics of Change

- *Aufgabe*

- Identifizierung potentieller Probleme (Inkonsistenzen, „resolution points“) bei Durchführung der Änderung
- Vorgabe vers. „evolution strategies“ zur Auflösung von Inkonsistenzen
 - Auswahl durch Nutzer
 - Automatische Entscheidung auf Basis einer globalen Strategie

- *Beispiel*

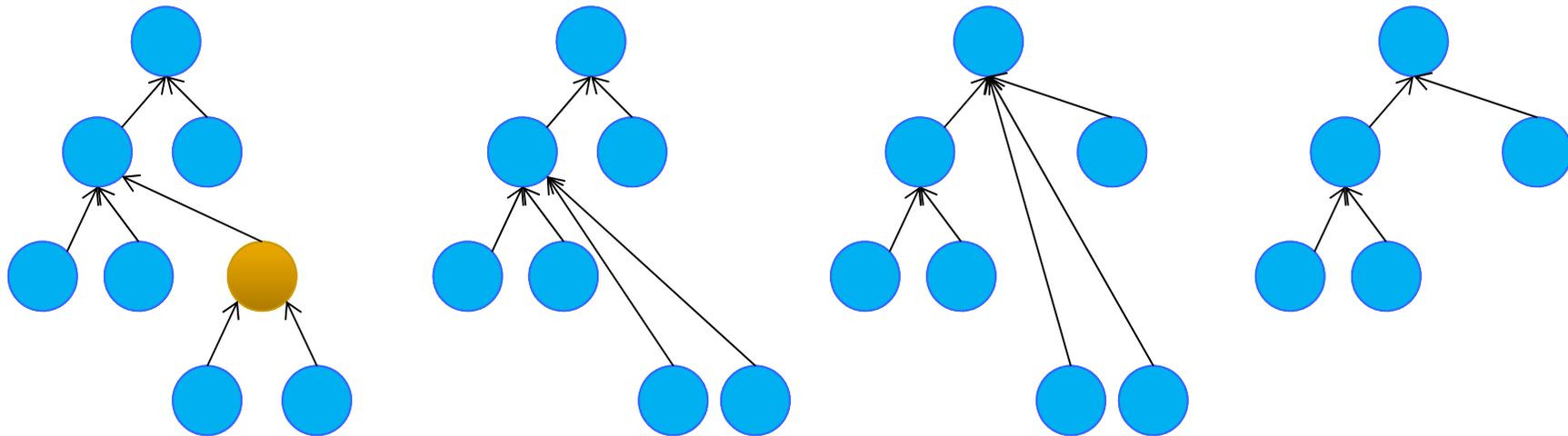
- *Delete_Concept(C)*

→ Wie wird mit „verwaisten“ Kindern (Subkonzepten) von *C* umgegangen?



3. Semantics of Change (2)

- Evolution strategies für *Delete_Concept(C)*



1. Einhängen unter Elternkonzept
2. Einhängen unter Wurzelkonzept
3. Ebenfalls Löschung der Kinder



4. Implementation

- *Aufgabe*

- Implementierung (Umsetzung) der aus den vorherigen Phasen angesammelten Änderungen
- Logging der umgesetzten Änderungen

- *Beispiel*

- Umsetzung von *Delete_Concept(C)*
- Je nach evolution strategy weitere Änderungen
 1. *Reconnect_to_parent(C)*
 2. *Reconnect_to_root(C)*
 3. *Delete_Subconcepts(C)*



5. Propagation / 6. Validation

- *Aufgabe - Propagation*
 - Propagierung der Änderungen in abhängige Strukturen, z.B. andere Ontologien die veränderte Ontologie verwenden
 - Annahme: Abhängigkeiten sind bekannt → rekursive Anwendung des Evolutionsprozesses in abhängigen Ontologien
- *Aufgabe - Validation*
 - Validierung des Ergebnis der umgesetzten Änderungen durch Ontologiedesigner
 - Option zum Undo von Änderungen
 - Initialisierung eines neuen Evolutionsprozesses



COnto-Diff (Complex Ontology Diff)

■ Probleme

- Einfache Änderungen oft nicht ausreichend
- Größe der Ontologien → kompakter Diff nötig
- Freiheiten bzgl. Modellierung, z.B. obsoletere

■ Ziel

- Bestimmung eines vollständigen, ausdrucksstarken und invertierbaren Evolution-Mapping zwischen zwei Ontologieversionen O_{old} und O_{new}

■ Vorschlag

- Regelbasierter DIFF-Ansatz
- Matching als Basis für Differenzbestimmung

Hartung, M., Groß, A., Rahm, E.: COnto-Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. Journal of Biomedical Informatics, 2012



Änderungsoperationen

■ Einfach (Basis)

- *add, del, map* für Konzepte C , Attribute A und Relationen R
 - *mapC(c1,c2)*: $c1$ wird durch $c2$ ersetzt

■ Komplex

- Nicht mengenwertig
 - *substitute, toObsolete, move, addLeaf, delLeaf...*
- Mengenwertig
 - *merge, split, addSubGraph, delSubGraph, ...*

■ Jede Operation ist invertierbar

- $merge(\{c1,c2,c3\}, c3) \leftrightarrow split(c3, \{c1,c2,c3\})$

■ Jede komplexe Operation kann durch eine Menge einfacher Operationen umgesetzt werden

- $merge(\{c1,c2,c3\}, c3)$
→ $mapC(c1,c3), mapC(c2,c3), mapC(c3,c3)$

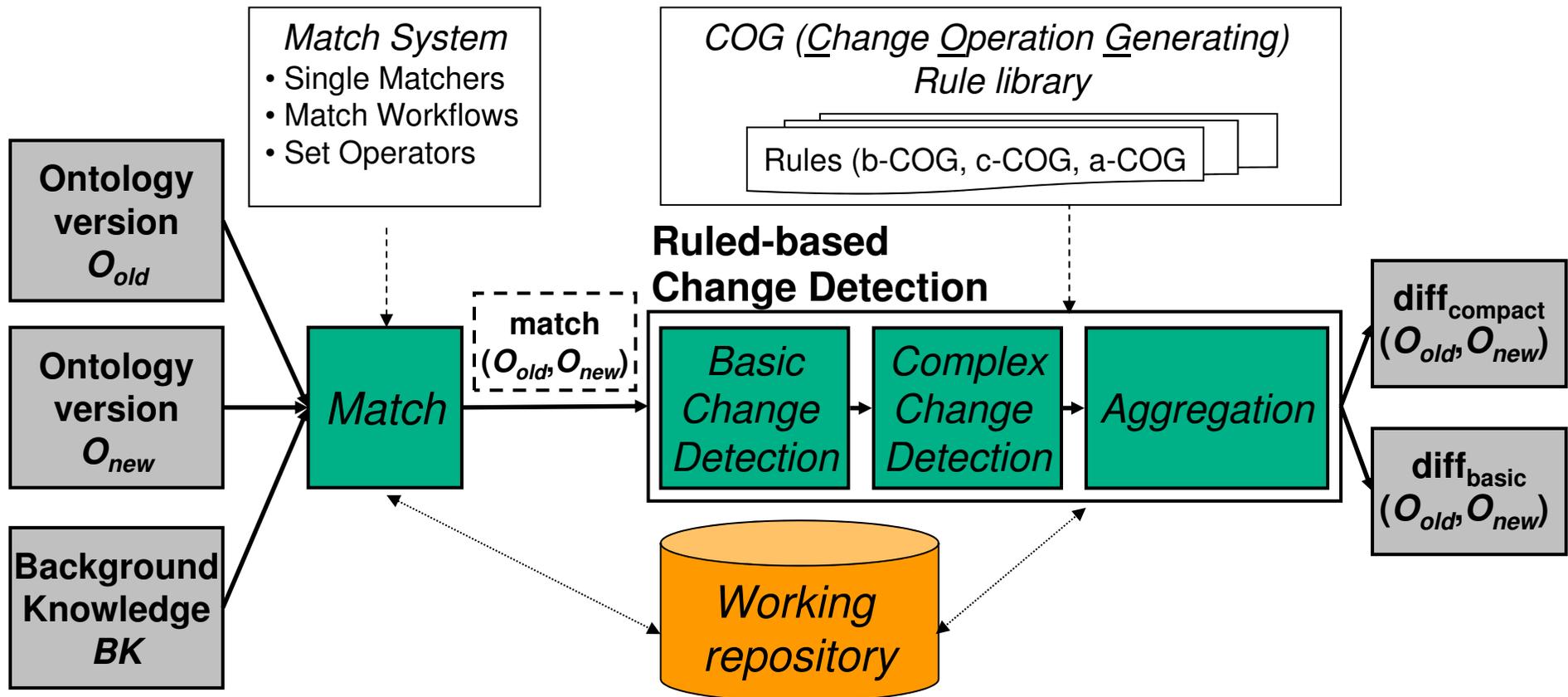


Match- und Evolution-Mappings

- Mapping-basierter Ansatz
- Match-Mapping $match(O_{old}, O_{new})$
 - Verbindet ähnliche (gleiche) Konzepte aus O_{old} und O_{new} miteinander (siehe Kapitel 5)
 - $matchC(c1, c2)$
- Evolution-Mapping $diff(O_{old}, O_{new})$
 - Änderungen zwischen O_{old} und O_{new} in Form einer Menge von Änderungsoperationen
 - Unterscheidung zwischen $diff_{basic}$ und $diff_{compact}$
 - $diff_{basic}$: nur einfache Änderungsoperationen
 - $diff_{compact}$: „so ausdrucksstark (kompakt) wie möglich“



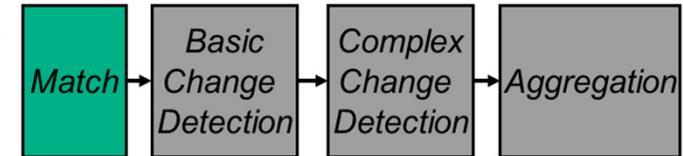
COntoDiff - Schematischer Überblick



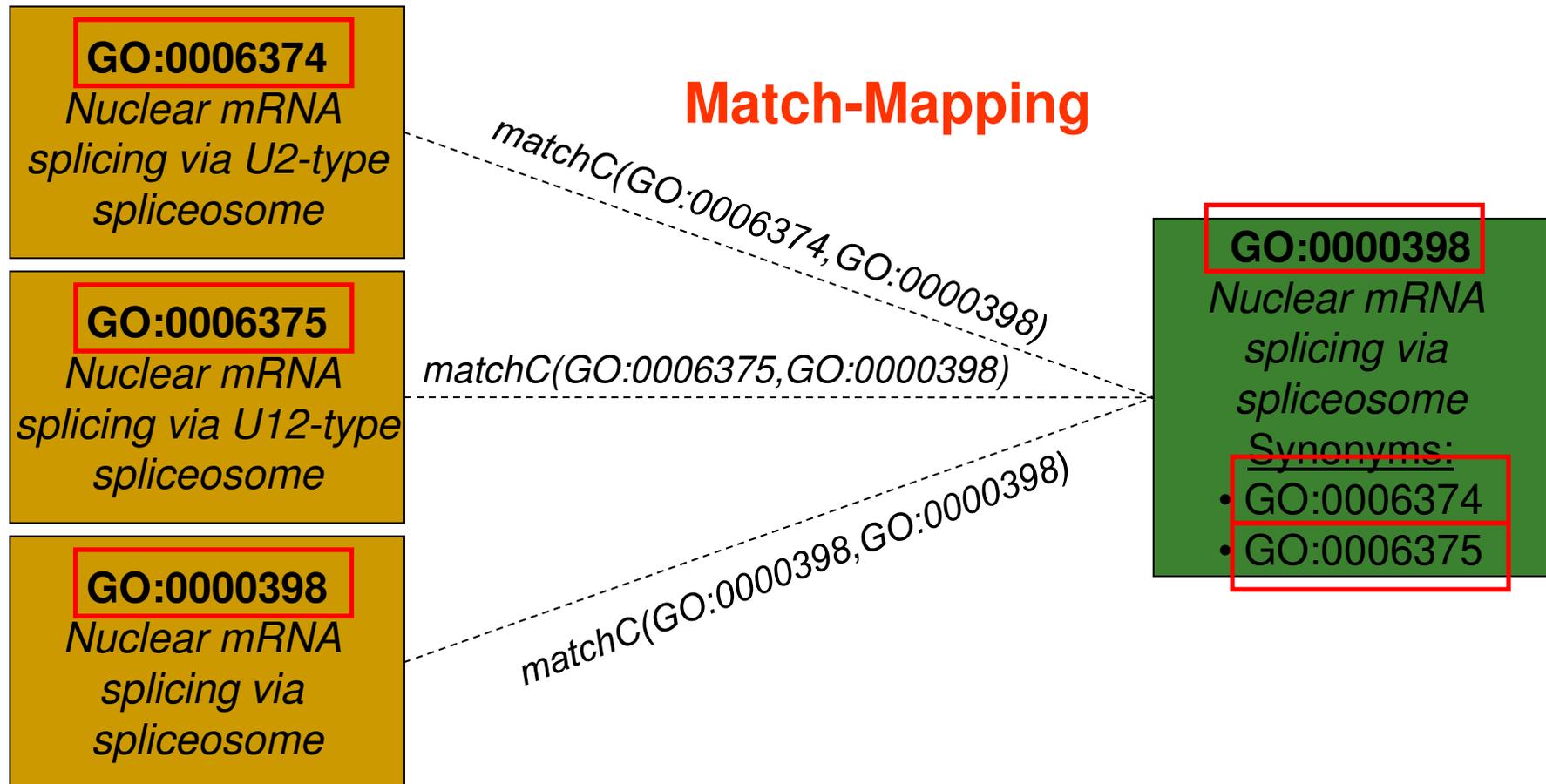
Beispiel: „Zusammenfassen“ (*merge*) mehrerer Konzepte zwischen 2008-01 und 2008-12 in GO Biologische Prozesse



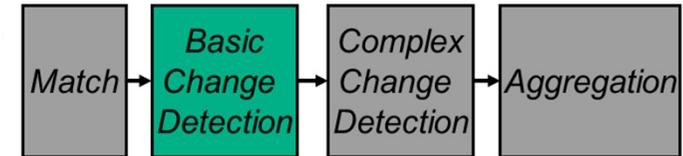
Match Phase



- *accession*-basiertes Matching



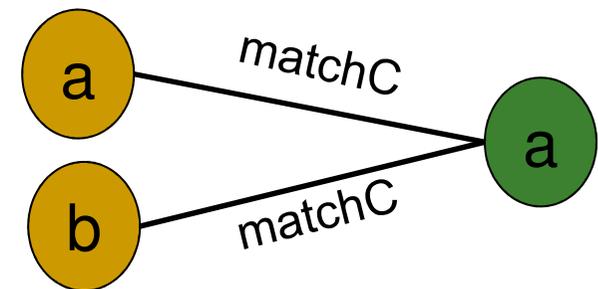
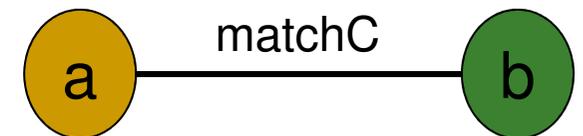
Basic Change Detection



- **b-COG** Regeln (COG: Change Operation Generating)
 - Erzeugung aller einfachen Änderungen
- **Eingabe:** Ontologieversionen O_{old}/O_{new} , Match-Mapping $match(O_{old}, O_{new})$, Regelmenge R_{b-COG}
- **Ergebnis:** $diff_{basic}(O_{old}, O_{new})$
- **merge:**

$(b_3): a \in O_{old} \wedge b \in O_{new} \wedge a \neq b \wedge matchC(a, b)$
 $\rightarrow \mathbf{create}[mapC(a, b)]$

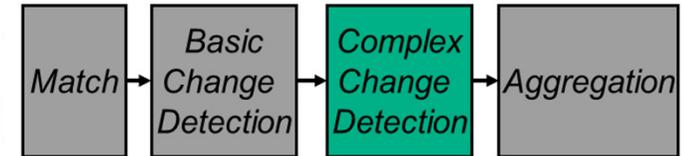
$(b_5): a \in O_{old}, O_{new} \wedge matchC(a, a)$
 $\wedge \exists b (b \in O_{old} \wedge matchC(b, a) \wedge a \neq b)$
 $\rightarrow \mathbf{create}[mapC(a, a)]$



$mapC(GO:0006374, GO:0000398)$ über (b_3)
 $mapC(GO:0006375, GO:0000398)$ über (b_3)
 $mapC(GO:0000398, GO:0000398)$ über (b_5)



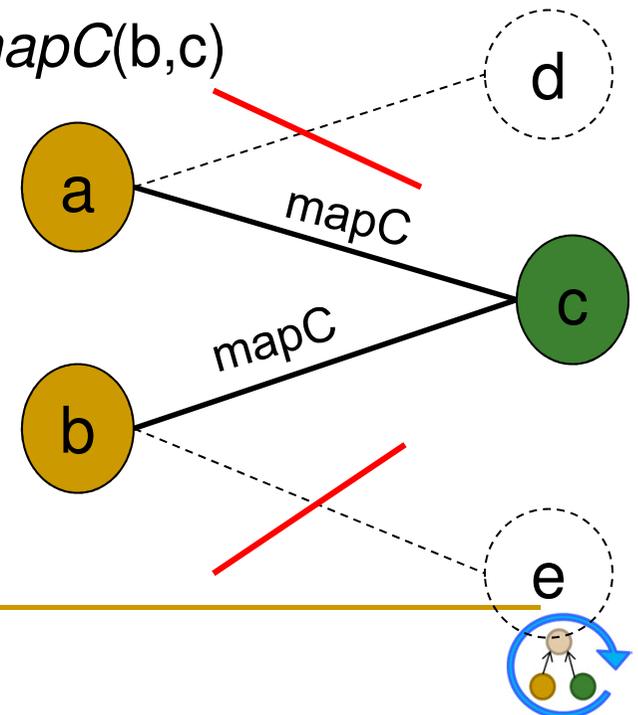
Complex Change Detection



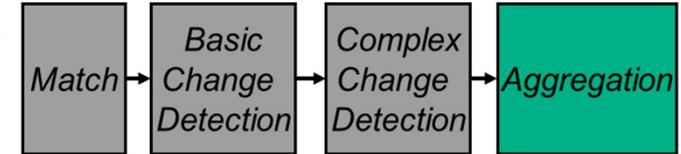
- **c-COG** Regeln
 - Erzeugung aller komplexen Änderungen
- **Eingabe:** Evolution-Mapping $diff_{basic}(O_{old}, O_{new})$,
Regelmenge R_{c-COG}
- **Ergebnis:** $diff(O_{old}, O_{new})$
- **merge:**

$a, b \in O_{old} \wedge c \in O_{new} \wedge a \neq b \wedge mapC(a, c) \wedge mapC(b, c)$
 $\wedge \nexists d (d \in O_{new} \wedge mapC(a, d) \wedge c \neq d)$
 $\wedge \nexists e (e \in O_{new} \wedge mapC(b, e) \wedge c \neq e)$
 $\rightarrow \mathbf{create}[merge(\{a\}, c), merge(\{b\}, c)],$
 $\mathbf{eliminate}[mapC(a, c), mapC(b, c)]$

$merge(\{GO:0006374\}, GO:0000398)$
 $merge(\{GO:0006375\}, GO:0000398)$
 $merge(\{GO:0000398\}, GO:0000398)$



Aggregation



- **a-COG** Regeln (rekursiv anwendbar)
 - Erzeugung mengenwertiger, komplexer Änderungen
- **Eingabe:** Evolution-Mapping $diff(O_{old}, O_{new})$, Regelmenge R_{a-COG}
- **Ergebnis:** $diff_{compact}(O_{old}, O_{new})$
- **merge:**

$c \in O_{new} \wedge A, B \subseteq O_{old} \wedge merge(A, c) \wedge merge(B, c) \wedge A \neq B$
→ **create**[$merge(A \cup B, c)$], **eliminate**[$merge(A, c), merge(B, c)$]



Evaluierung

- Gene Ontology (GO)
- Zwei Versionsvergleiche
 - 2008: 2008-01 → 2009-01
 - 2009: 2009-01 → 2010-01



$O_{old} - O_{new}$	$ O_{old} $ (C , R)	$ O_{new} $ (C , R)	match	diff _{compact}
GO₂₀₀₈₋₀₁ - GO₂₀₀₉₋₀₁	66.121 ⌞ 25.774 ⌞ 40.347	75.180 ⌞ 27.870 ⌞ 47.310	25.774	8.450
GO₂₀₀₉₋₀₁ - GO₂₀₁₀₋₀₁	75.180 ⌞ 27.870 ⌞ 47.310	84.714 ⌞ 30.751 ⌞ 53.963	27.870	4.284

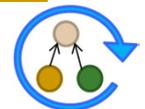


Evaluierung

- Details zu $diff_{compact}$ und $diff_{basic}$

	GO	
	2008	2009
$ diff_{basic} $	15.781	13.504
$ diff_{compact} $	8.450	4.284
\llcorner #basic	5.594	1.671
\llcorner #complex	2.856	2.613
ratio in %	53,5%	31,7%

	GO	
	2008	2009
add	4.187	1.355
del	1.407	316
map	0	0
addLeaf	768	796
delLeaf	0	0
merge	70	83
move	1.499	1.200
substitute	0	1
toObsolete	225	66
addSubGraph	294	467
delSubGraph	0	0
Σ	8.450	4.284



Speichereffiziente Versionierung

■ Problem

- ❑ Sehr große Ontologien (bis zu 100.000 Konzepten)
- ❑ Versionen oft nur wenig Tage / Monate gültig

■ Speichereffiziente Versionierung erforderlich

- ❑ Zugang zu verschiedenen Ontologieversionen
- ❑ Basis für Evolutionsanalysen (OnEX, ...)

■ Vorschlag

- ❑ Versionierung mittels Zeitstempeln für Ontologeelemente
- ❑ Über Zeitstempel können Ontologieversionen für jeden beliebigen Zeitpunkt rekonstruiert werden
- ❑ Vergabe / Änderung von Zeitstempeln bei Integration einer neuen Ontologieversion

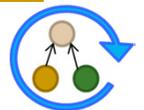
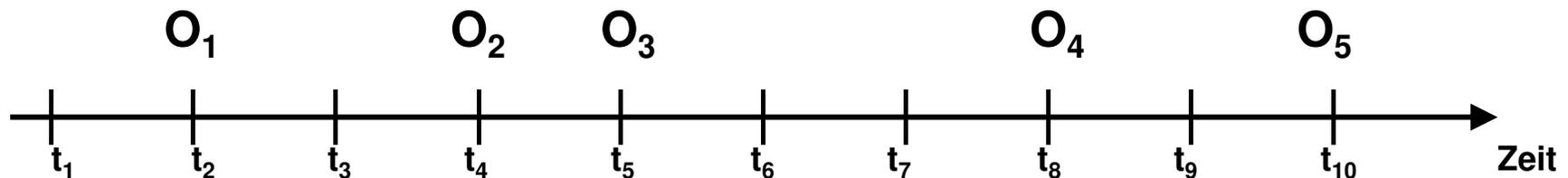
Kirsten, T., Hartung, M., Groß, A., Rahm, E.: Efficient Management of Biomedical Ontology Versions. In Proc. 4th Intl. Workshop on Ontology Content (Part of the OTM Conferences & Workshops), 2009



Lineares Versionierungsmodell

■ Lineares Versionierungsschema

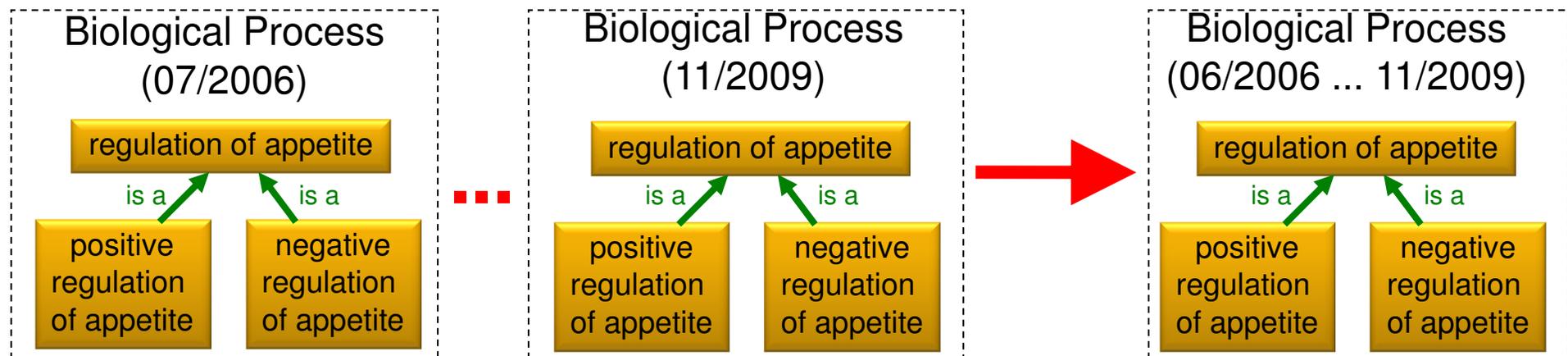
- Kette von Ontologieversionen
- Eine Ontologieversion O_j hat exakt eine Vor- bzw. Nachfolgeversion O_{j-1} bzw. O_{j+1}
- Erste / letzte Version bilden Ausnahmen



Versionierung über Zeitstempel

■ Beobachtungen

- Ontologiekonzepte, deren Attribute sowie Beziehungen zwischen Konzepten sind innerhalb eines bestimmten Zeitraums gültig
- Eine Ontologieversion repräsentiert den Zustand (Snapshot) einer Ontologie zu einem konkretem Zeitpunkt t



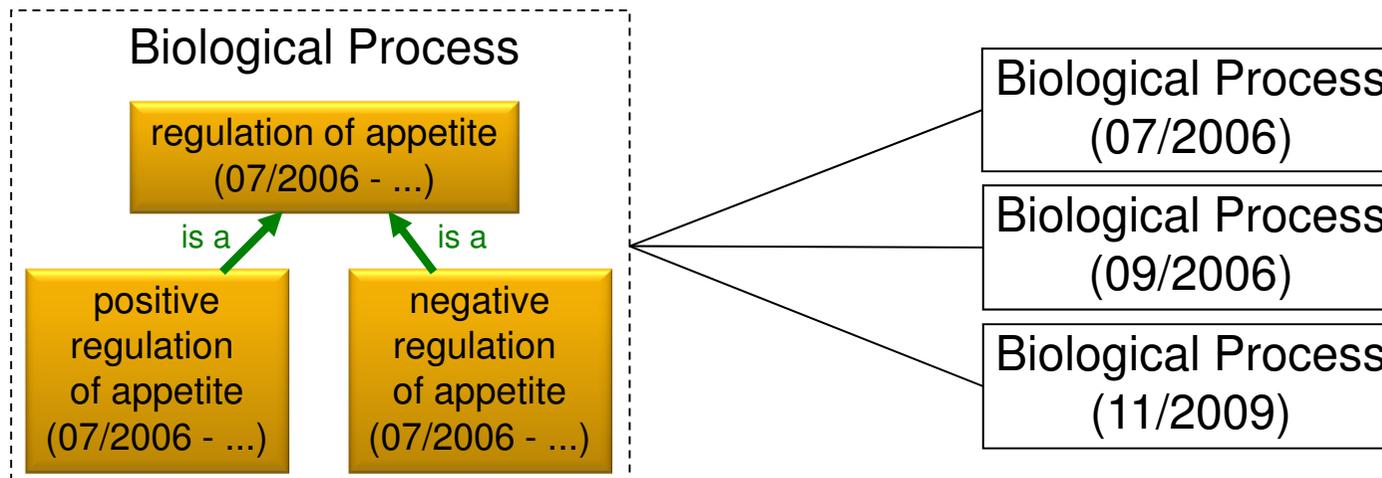
- **Idee:** Speichere Konzepte (Attribute, Beziehungen) **nur einmal** und weise einen **Gültigkeitszeitraum** zu



Versionierung über Zeitstempel (2)

■ Zuordnung einer Lebenszeit (Gültigkeitszeitraum)

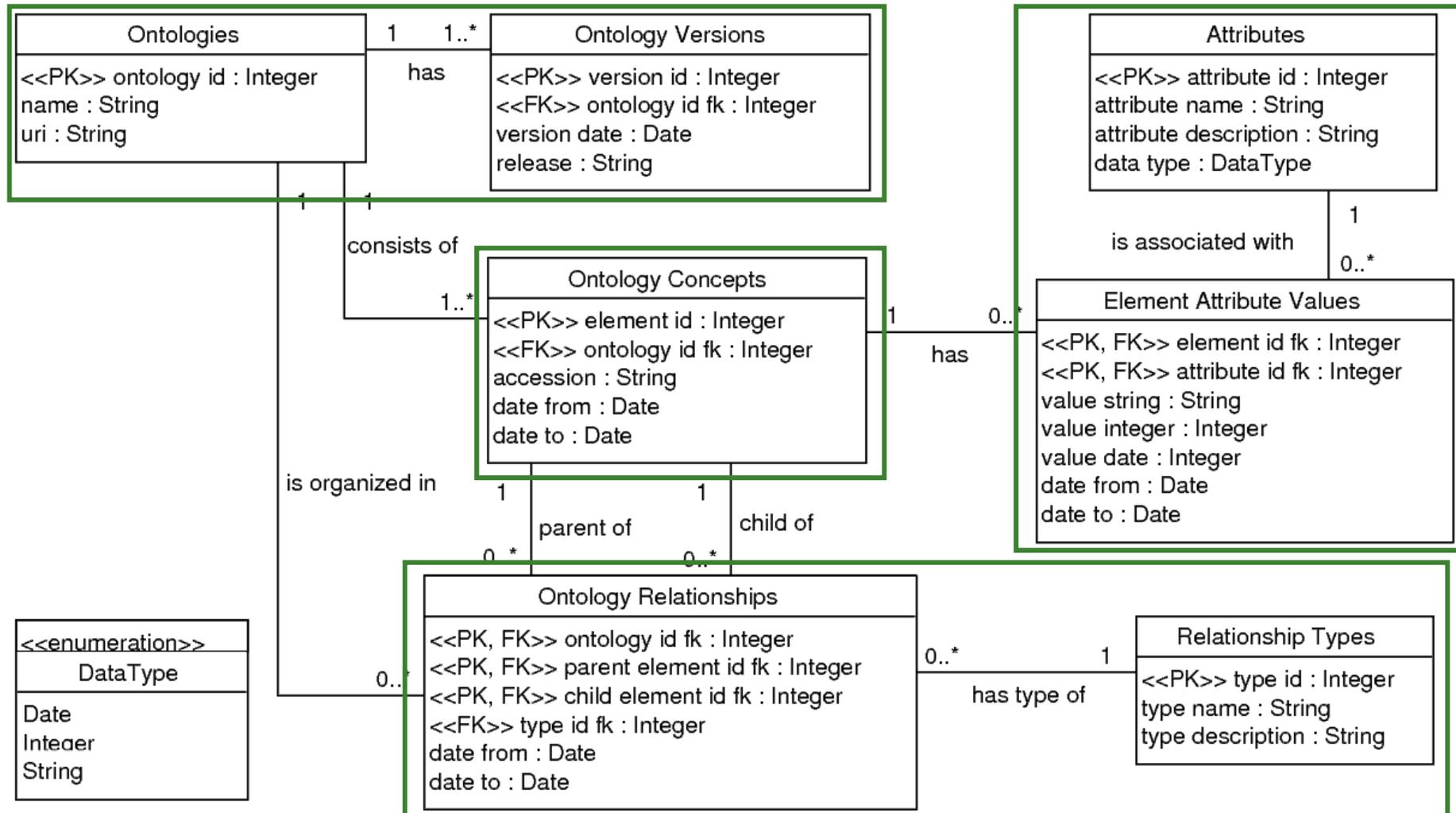
- Jedes Element bekommt eine Lebenszeit zugeordnet
- t_{start} – Zeitpunkt des ersten Auftretens
- t_{end} – Zeitpunkt des letzten Auftretens
- Gültige Elemente zum Zeitpunkt t : $t_{\text{start}} \leq t \leq t_{\text{end}}$



- ### ■ Rekonstruktion einer Ontologieversionen über t_{start} und t_{end} sowie dem Veröffentlichungsdatum der Version



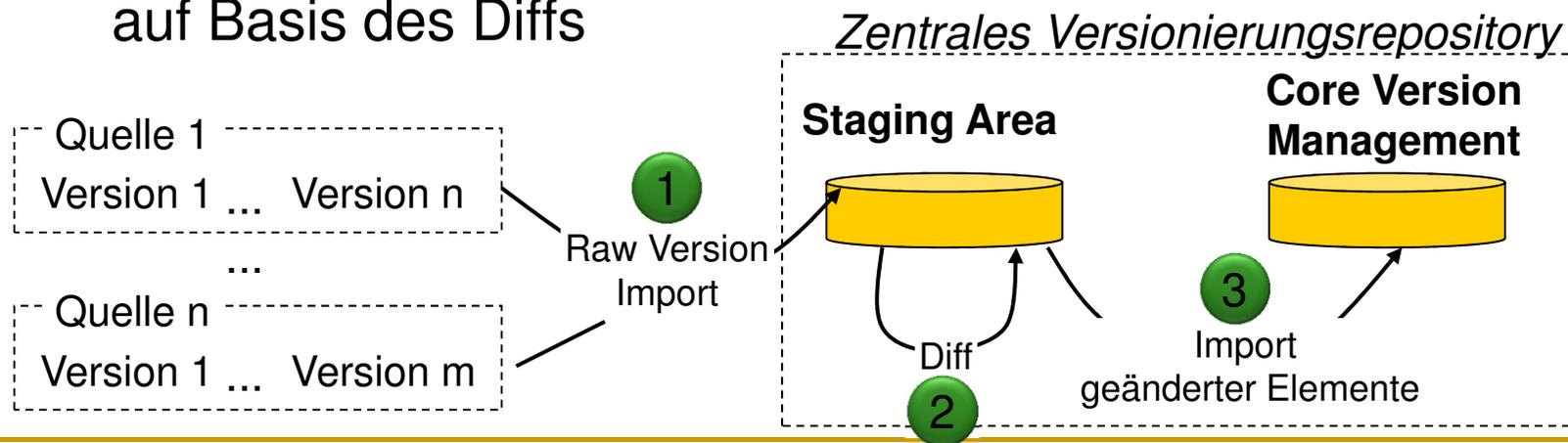
Schema für speichereffiziente Versionierung



Integration von Versionen

■ Dreiphasiger Import

- 1 Download und Integration der Originalversionen in eine Staging Area
 - Quellen: SVN, CVS, FTP, Webseiten, ...
 - Formate: OWL, OBO, XML, CSV, ...
- 2 Berechnung des Diff mit der letzten aktuellen Version
 - *add / del* von Konzepten, Attributen, Relationships
- 3 Integration neuer Elemente und Anpassung der Zeitstempel auf Basis des Diffs



Integration von Versionen am Beispiel

- **Beispiel: Integration der GO-CC Version von 2007-06**
 - **Löschung von GO:0009572**
 - **Einfügung von GO:0000446**
 - **Namensänderung von GO:0009356**

<i>Concepts</i>		
<i>accession number</i>	<i>start</i>	<i>end</i>
GO:0009572	2002-02	2007-05
GO:0000446	2007-06	
...		

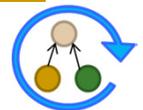
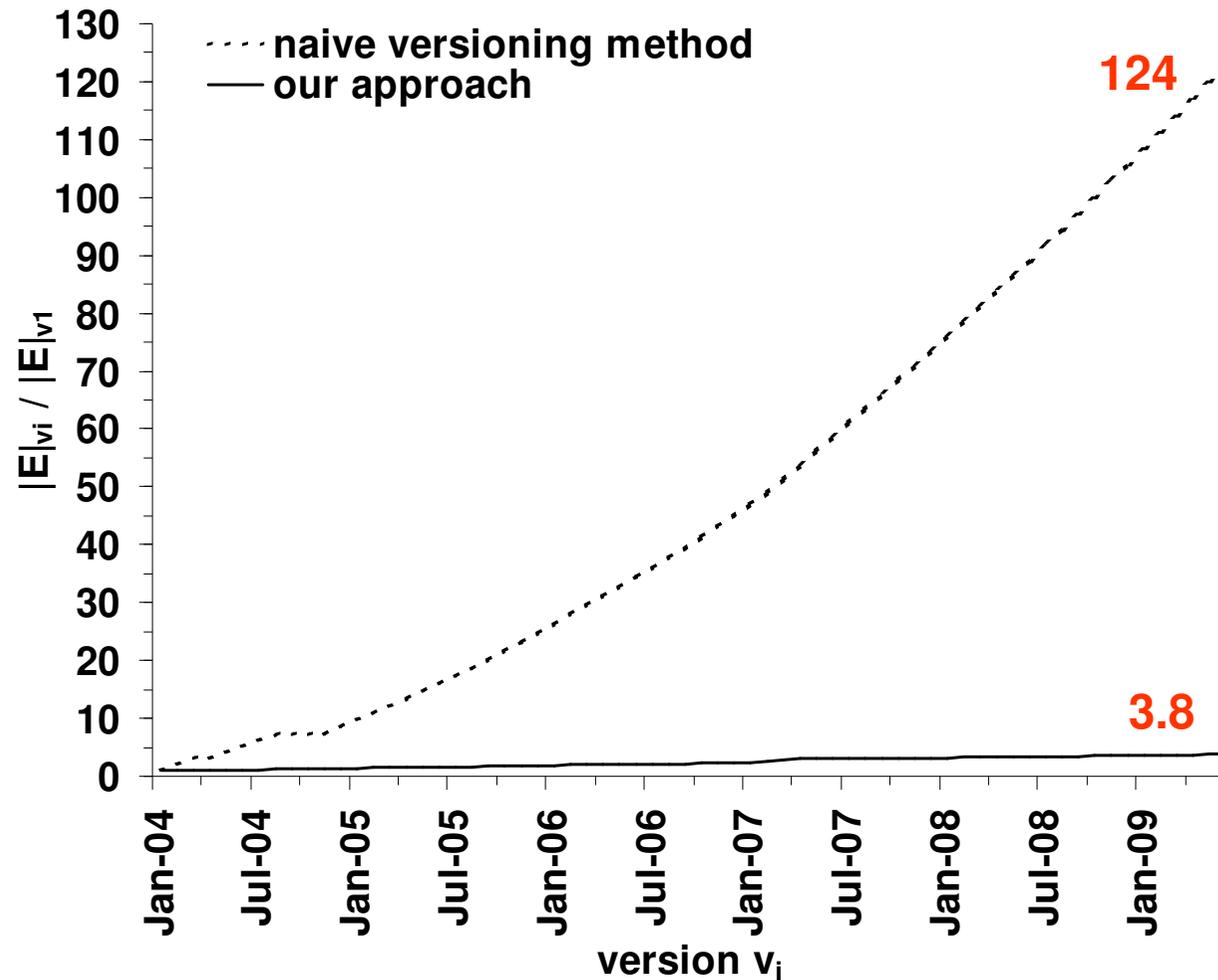
<i>Relationships</i>				
<i>source</i>	<i>target</i>	<i>type</i>	<i>start</i>	<i>end</i>
GO:0009572	GO:0044459	is_a	2006-05	2007-05
GO:0009572	GO:0009510	part_of	2003-05	2007-05
GO:0000446	GO:0000347	is_a	2007-06	
GO:0000446	GO:0008023	is_a	2007-06	

<i>Attributes</i>				
<i>concept</i>	<i>attribute</i>	<i>value</i>	<i>start</i>	<i>end</i>
GO:0009572	name	desmotubule central rod	2002-02	2007-05
GO:0009572	obsolete	false	2002-02	2007-05
GO:0000446	name	nucleoplasmatic THO complex	2007-06	
GO:0000446	obsolete	false	2007-06	
GO:0000446	definition	The THO complex when ...	2007-06	
GO:0009356	name	p-aminobenzoate synthetase complex	2002-12	2007-05
GO:0009356	name	aminodeoxychorismate synthase complex	2007-06	



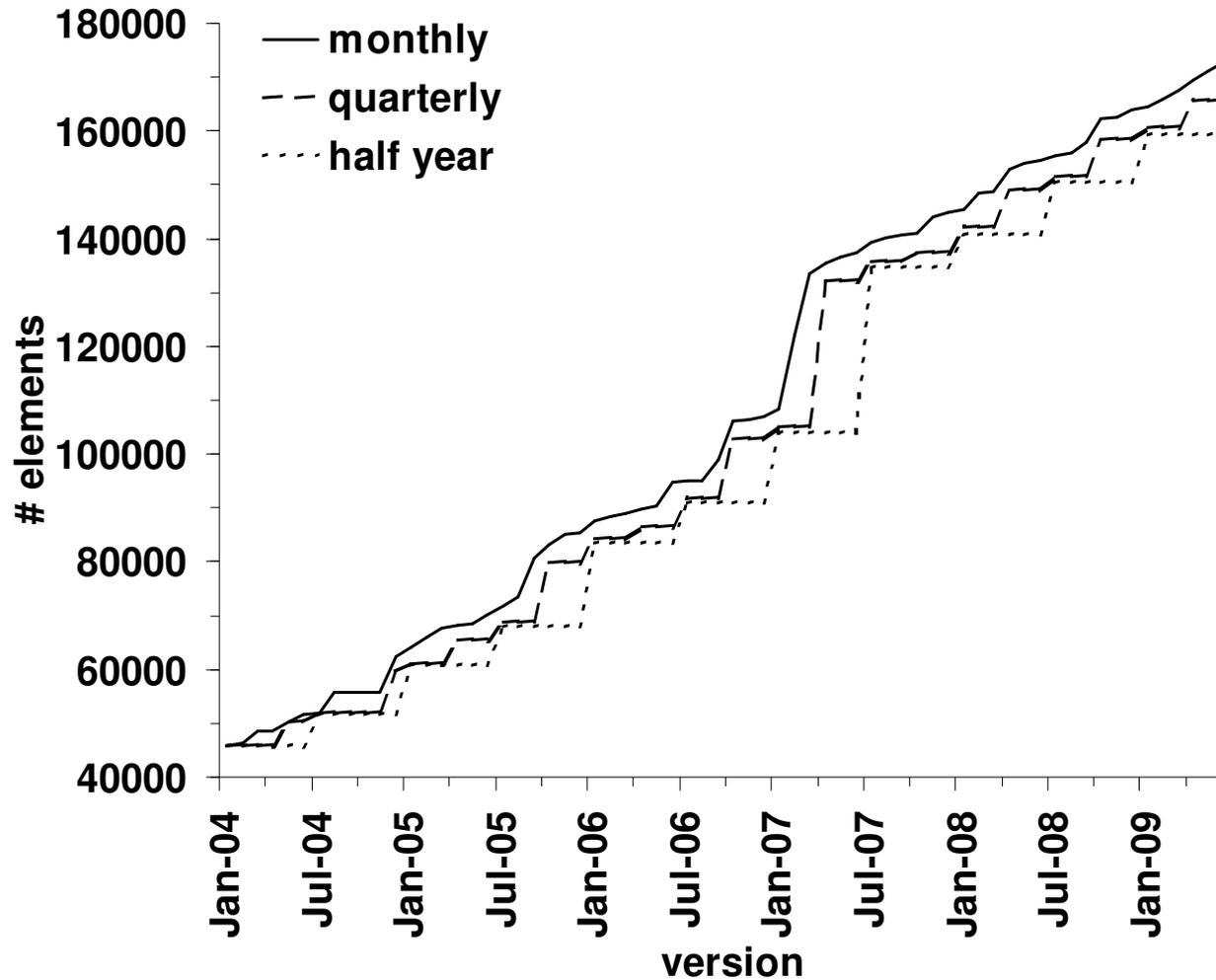
Evaluierung – Zeitstempel vs. Naiv

- 62 GO-BP Versionen zwischen 2004-01 und 2009-06



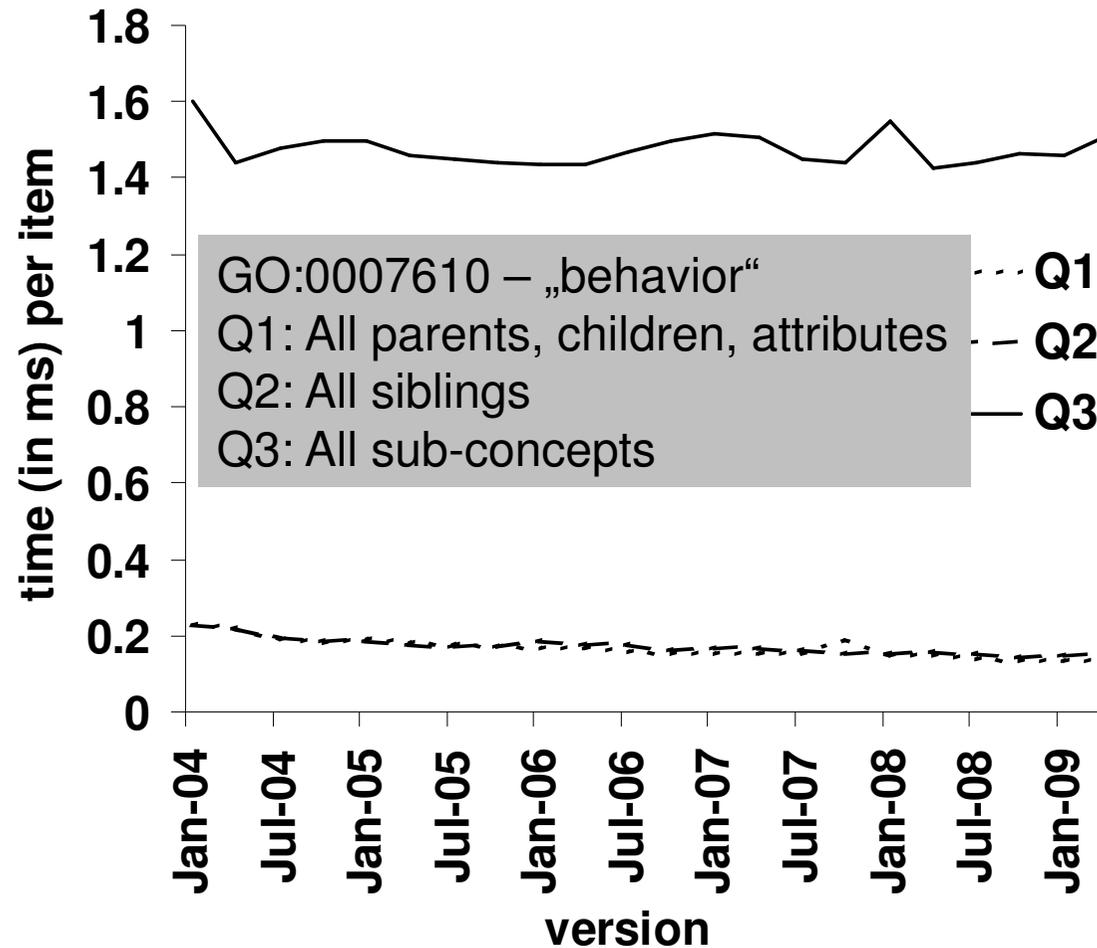
Evaluierung – Versionierungsintervalle

- 62 GO-BP Versionen zwischen 2004-01 und 2009-06



Evaluierung – Effizienz von Anfragen

- 62 GO-BP Versionen zwischen 2004-01 und 2009-06



Zusammenfassung

- **Dynamik in Ontologien**

- Ontologien unterliegen ständigen Änderungen
- Vers. Gründe für Anpassungen (Evolution)
- Probleme durch Dynamik
 - Anpassung abhängiger Daten
 - Einfluss auf Applikationen und Analysen

- **Drei Aspekte der Dynamik von Ontologien**

- Prozess-orientierte Anpassung von Ontologien
- Berechnung eines Diff zwischen Ontologieversionen
- Speichereffiziente Versionierung von Ontologien

