

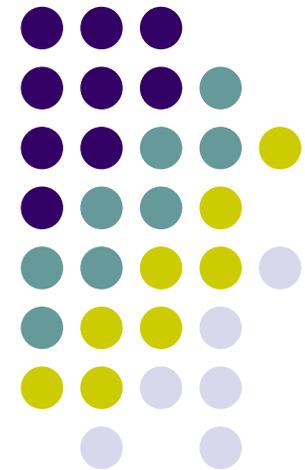
## **Diplomarbeit:**

GOMMA: Eine Plattform zur flexiblen Verwaltung und Analyse von Ontologie Mappings in der Bio-/Medizininformatik

---

Bearbeiter: Shuangqing He

Betreuer: Toralf Kirsten, Michael Hartung



**Universität Leipzig**  
**Institut für Informatik**  
**- Abteilung Datenbanken -**

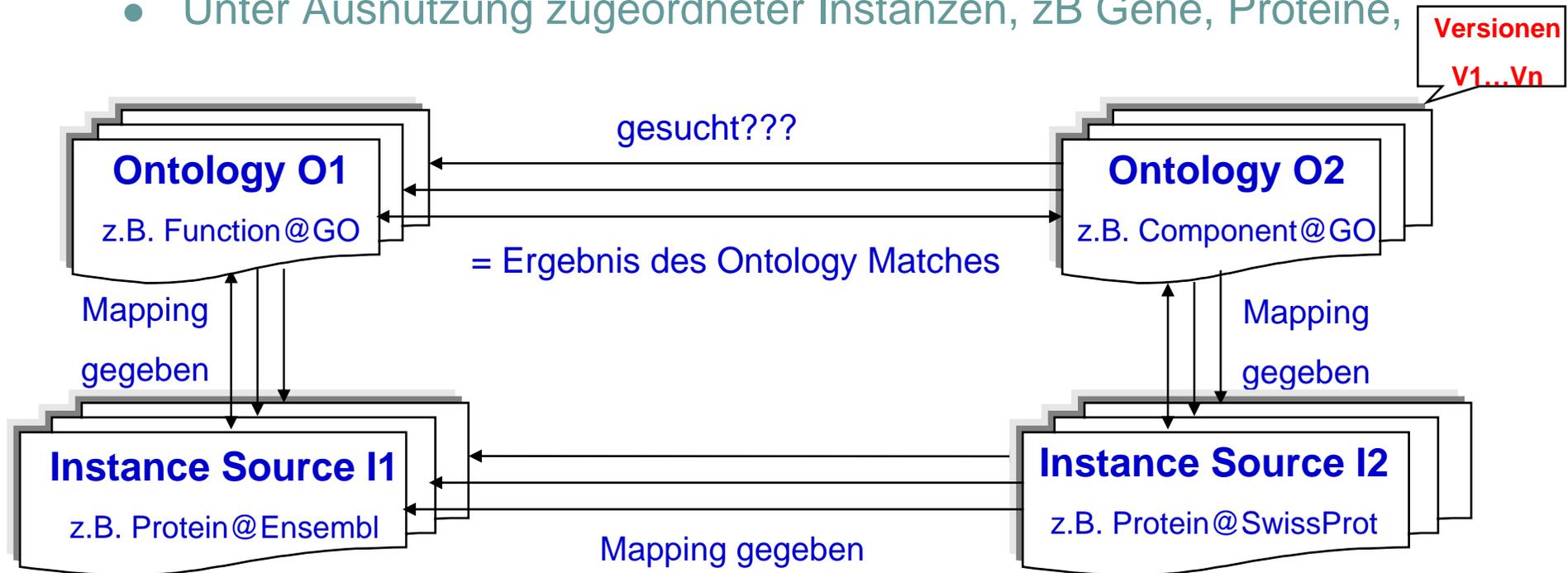
# Agenda

- Motivation
- Nutzungsszenarien
- Systemarchitektur und -funktionalitäten
- Implementierung
- Zusammenfassung & Ausblick



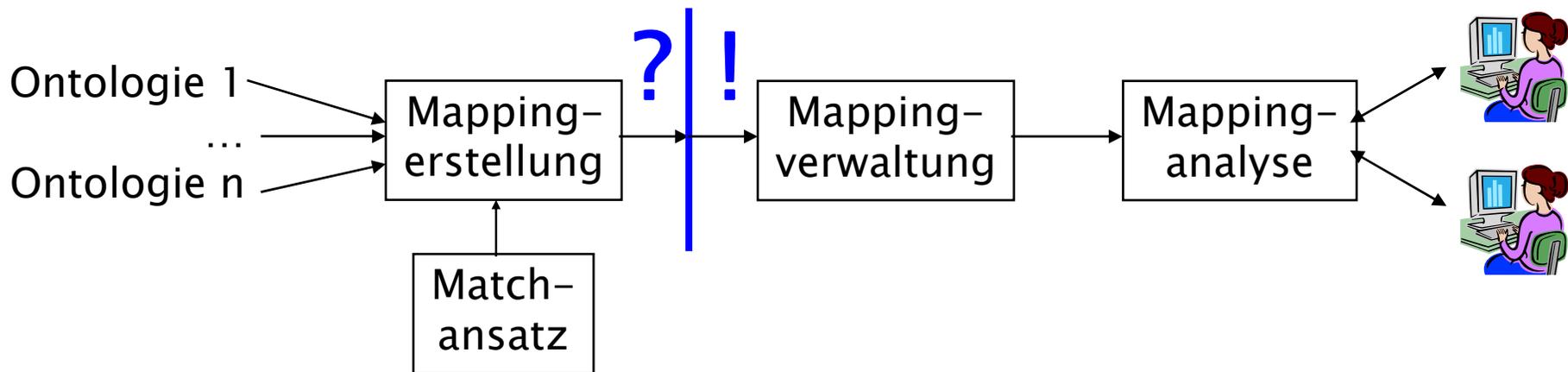
# 1. Motivation

- Viele Quellen und Ontologien in den Bereichen Bio- und Medizininformatik, z.B. GeneOntology, UMLS, ...
- Verschiede Ansätze für das Ontology Matching (=Erstellung eines Mappings) vorhanden
  - Basierend auf Ontologie-Metadaten
  - Unter Ausnutzung zugeordneter Instanzen, zB Gene, Proteine,



# 1. Motivation

- Problemstellung: Mapping ist erstellt → Was nun?



- Zielstellung: Entwicklung eines webbasierten Systems zur flexiblen Verwaltung und Analyse
  - diverser Ontologien, wobei jede Ontologie als Menge von Konzepten und deren Beziehungen repräsentiert werden kann
  - vorberechneter Mappings zwischen Ontologien

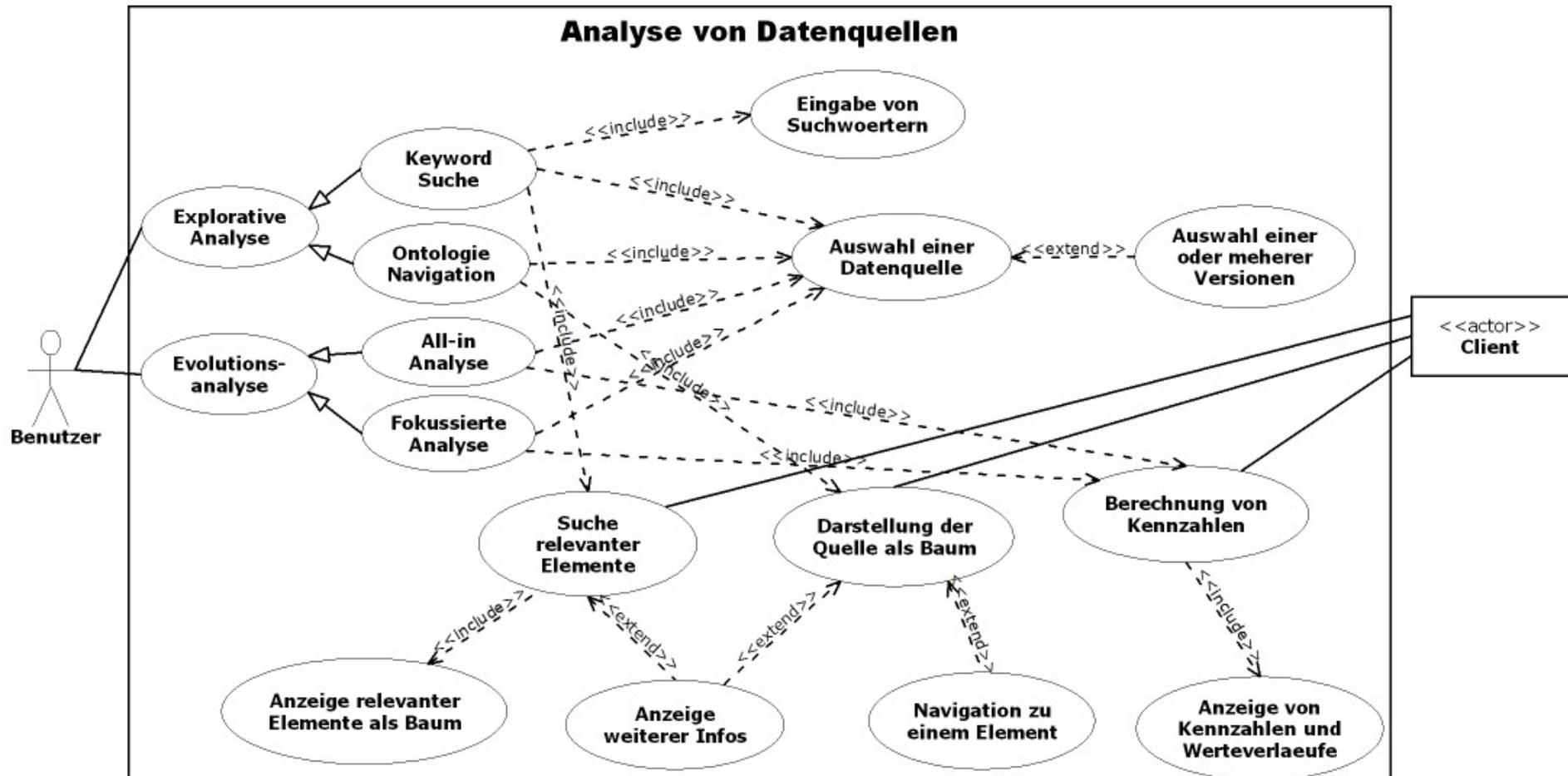
# 1. Motivation



- Analyseanwendungen:
  - Explorative Analyse bzw. Evolutionsanalyse von Datenquellen
  - Explorative Analyse bzw. Evolutionsanalyse von Mappings
  - Mapping-Generierung
    - Filterung
    - Mengenmanipulation mit union, intersect und majority
    - Differenzenermittlung
    - Komposition
- Neuer Prototyp: GOMMA-Kernel, eine vorhandene API, die die Zugriffe auf das Mapping-Repository steuert/ anbietet.

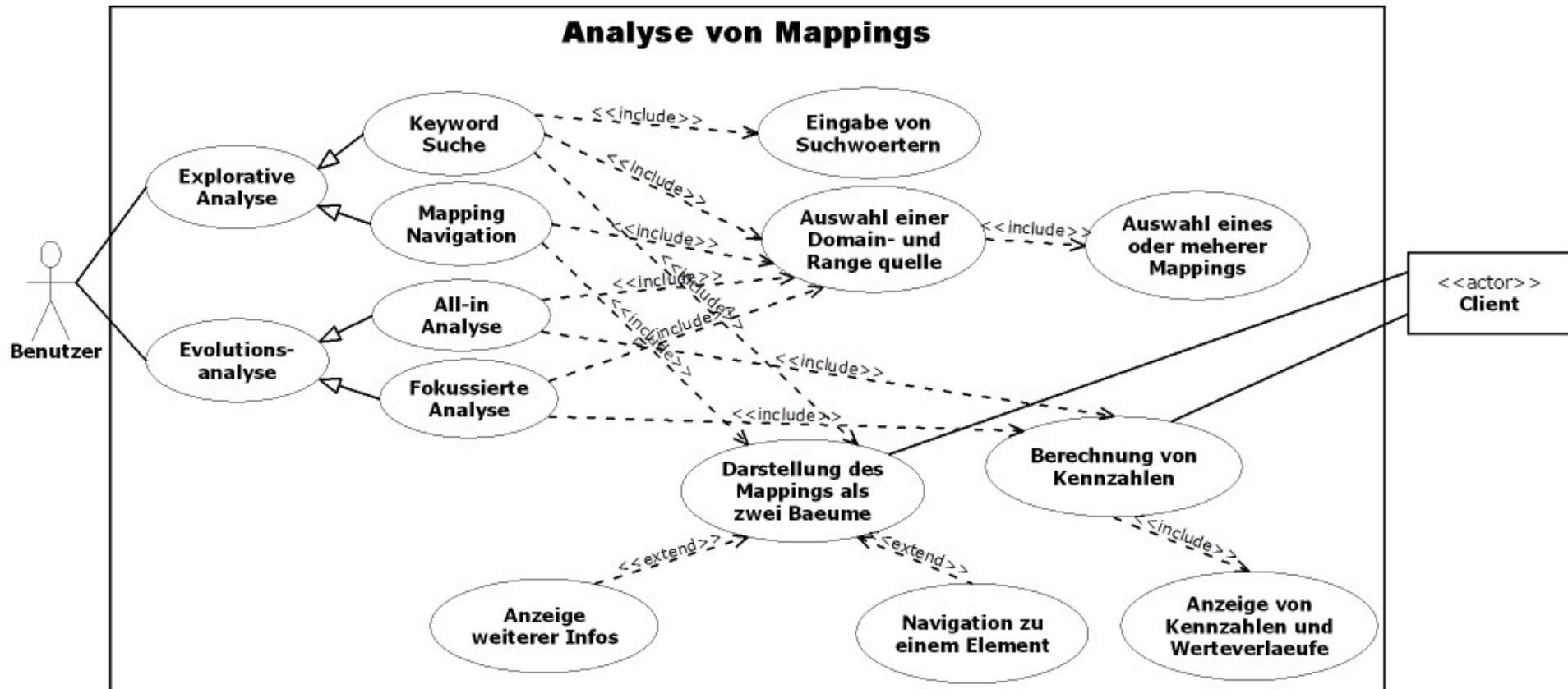
# 2. Nutzungsszenarien

- Analyse von Datenquellen



# 2. Nutzungsszenarien

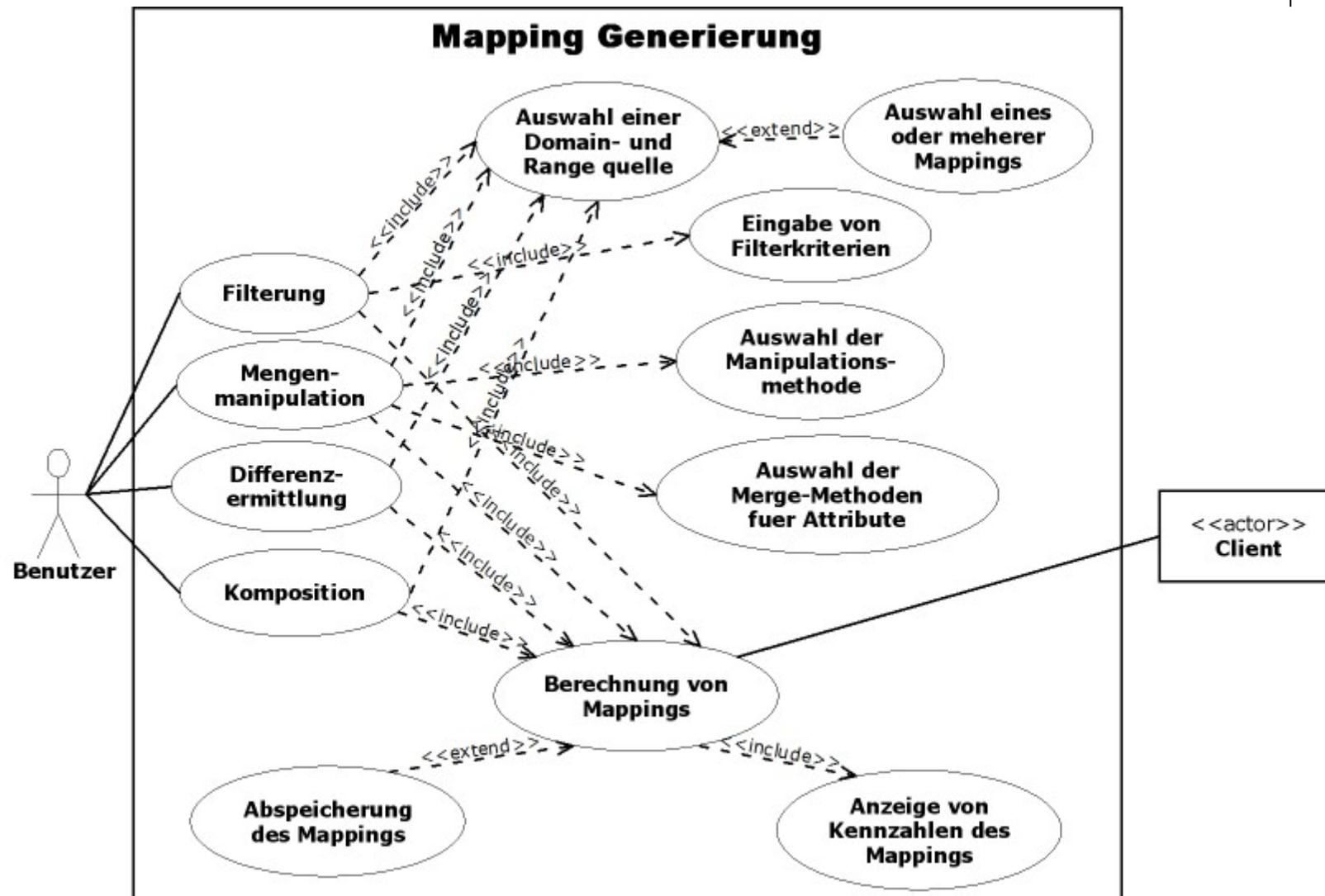
- Analyse von Mappings



# 2. Nutzungsszenarien



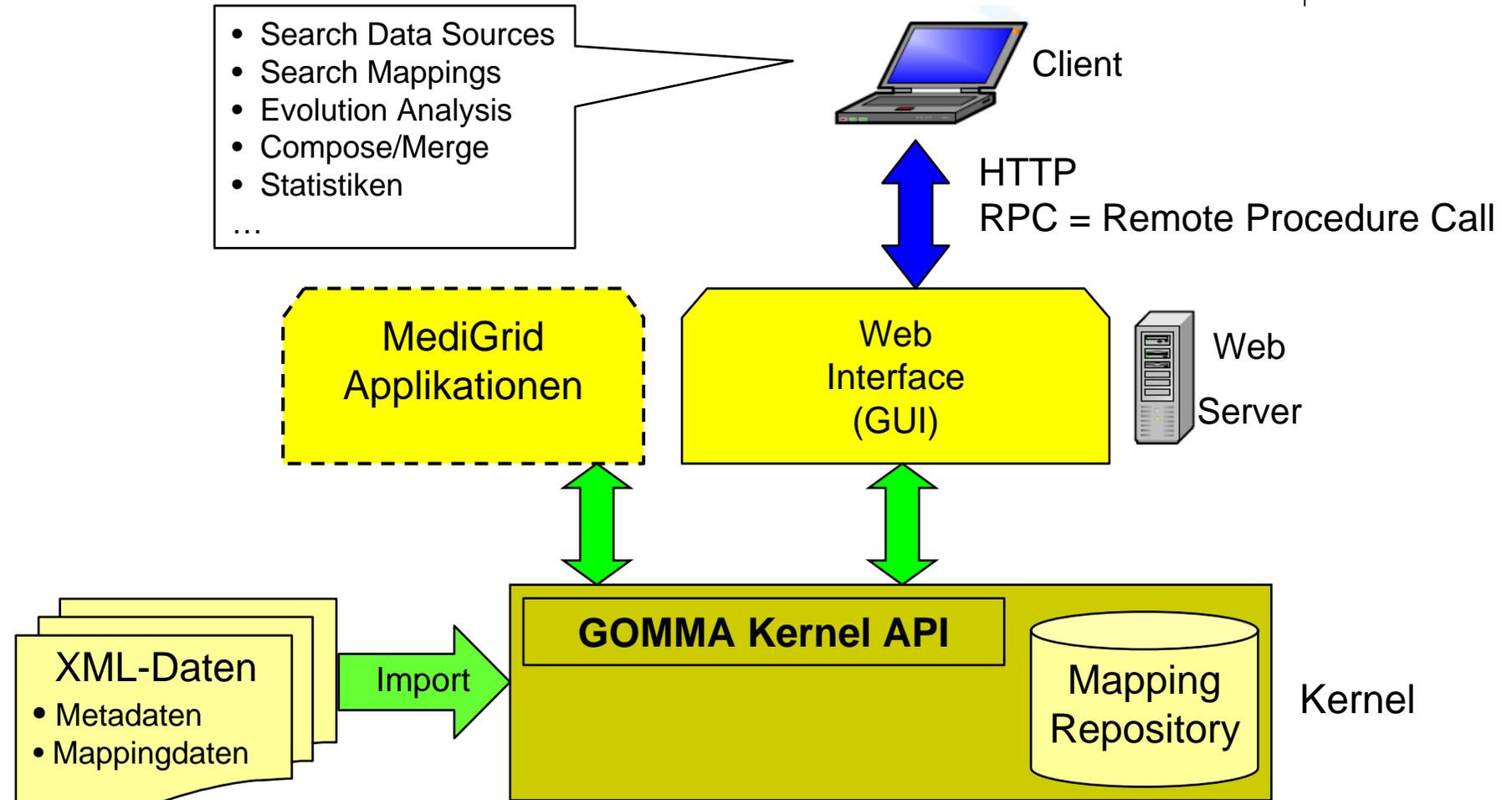
- Mapping Generierung



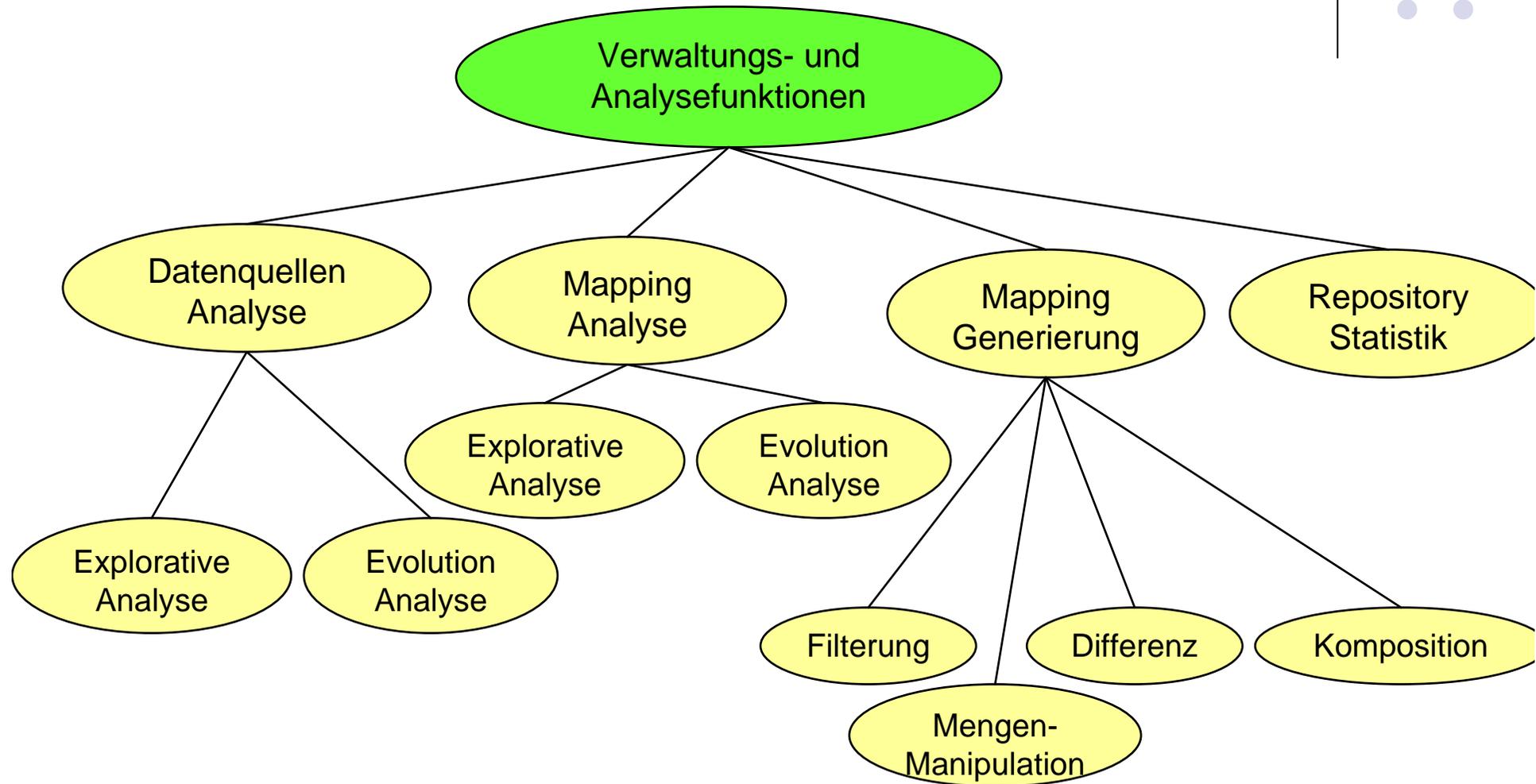
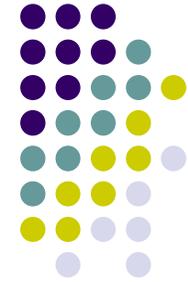
# 3. Systemarchitektur



Komponenten des Generic Ontology Mapping Management & Analysis



# 3. Systemfunktionalitäten im Überblick



# 4. Implementierung: Überblick

- Middleware: Kernel
- Relationale Datenbank: MySQL
- Tomcat Web-Server
- GUI: WEB 2.0 Erweiterungen
  - AJAX-Framework: Google Web Toolkit (GWT)
  - Google Chart API zur Visualisierung von Evolutionsanalyse



# 4. Implementierung: Was ist GWT ?

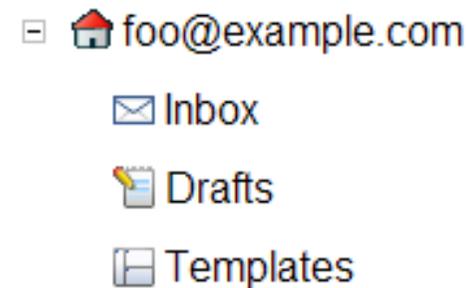


- GWT (Google Web Toolkit) :
  - GWT ist ein open-source Framework zur Entwicklung von Webanwendungen
  - Besonderheit: Java-nach-Javascript-Compiler
  - Die Kommunikation mit einem entfernten Server ist über **RPC**(Remote Procedure Calls) möglich
  - Widget-Paket zur Gestaltung der graphischen Oberfläche (ähnlich wie Swing)

- z.B. **ListBox**,



**Tree**,



## 4. Implementierung: Warum GWT ?



- Betrachtung anderer Frameworks:
  - Spring 2
    - bietet keine Widget-Library an, z.B. zur Baumdarstellung
  - Yahoo UI Library oder Ext JS
    - Beide sind JavaScript-Library => die Kommunikation mit dem Kernel via JSP oder Servlet notwendig
- Wesentliche Vorteile von GWT:
  - Der client-seitige Code kann komplett in Java erstellt werden => erhebliche Vorteile in der Entwicklung
  - Die Kommunikation mit dem Kernel über **RPC**(Remote Procedure Calls) ist viel einfacher.

# 4. Implementierung: GUI Startseite



Home	Data Sources	Mappings	Statistics	Help	Contact
Explorative Analysis ▶		Keyword Search	mappings between different ontologies. Please use the menu on the s underlying concepts:		
Evolution Analysis ▶		Ontology Navigation			

The GOMMA provides top to explore the full range of mappings between different ontologies. Please use the menu on the top to explore the full range of underlying concepts:

- **Home:** This page
- **Data Sources:** Explorative analysis and evolution analysis of data sources
- **Mappings:** Explorative analysis, evolution analysis and generation of mappings
- **Statistics:** Get statistics of available mappings and sources in GOMMA
- **Help:** Get help
- **Contact:** Feedback and contact information

**Current Version:**

- Program version: 0.5
- Last update: **05.06.2008**

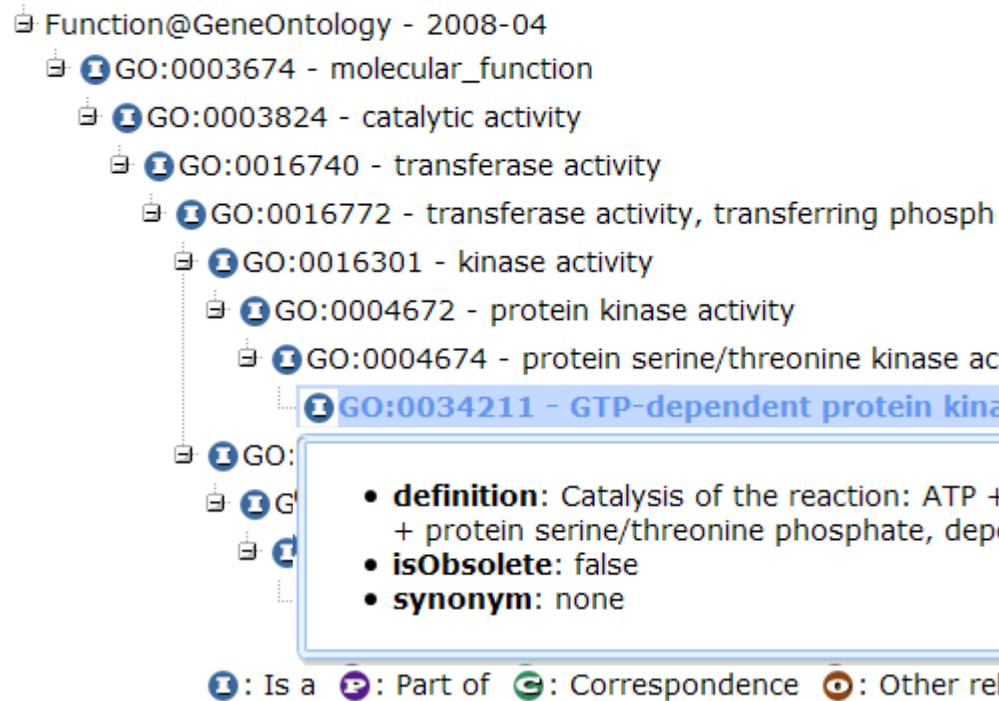
# 4. Explorative Analyse von Datenquellen



## • Keyword-Suche

Source:  Version:

Keyword:



1. Auswahl einer Datenquelle
2. Auswahl einer Version
3. Eingabe von Schlüsselwörtern
4. Suche relevanter Elemente der Datenquelle
5. Anzeige der relevanten Elemente als ein Baum

# 4. Anfrageformulierung

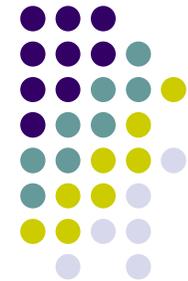
- Anfrageformulierung (autocomplete):

Keyword:

- GTP-dependent **protein** kinase activity - GO:0034211
- cyclic nucleotide-dependent **protein** kinase activity - GO:0004690
- acyl-[acyl-carrier-**protein**] desaturase activity - GO:0045300
- protein** self-association - GO:0043621
- transmembrane receptor **protein** tyrosine kinase adaptor protein activity - GO:0005068
- DNA-dependent **protein** kinase activity - GO:0004677
- protease** binding - GO:0002020
- pro**cytosylphosphatidylinositol phospholipase D activity - GO:0004621
- DNA polymerase **processivity** factor activity - GO:0030337
- alpha-1,3-mannosylglyco**protein** 4-beta-N-acetylglucosaminyltransferase activity - GO:0008454

Keyword:

- acetate transmembrane transporter activity - GO:0015**123**
- death receptor binding - GO:0005**123**
- glucanoyltransferase activity - GO:0042**123**
- N-acylmannosamine 1-dehydrogenase activity - GO:0050**123**
- cystathionine gamma-lyase activity - GO:0004**123**
- quinoline-4-carboxylate 2-oxidoreductase activity - GO:0047**123**
- cholesterol 7-alpha-monooxygenase activity - GO:0008**123**
- RAL GTPase activator activity - GO:0017**123**



# 4. Explorative Analyse von Datenquellen



## • Ontologie Navigation

Source:  Version:

Component@GeneOntology - 2008-04

- GO:0005575 - cellular\_component
  - GO:0045202 - synapse
  - GO:0044456 - synapse part
  - GO:0043226 - organelle
  - GO:0031974 - membrane-enclosed lumen
  - GO:0031012 - extracellular matrix**
    - definition:** A structure lying external to one or more cells, which provides structural support for cells or tissues; may be completely external to the cell (as in animals) or be part of the cell (as in plants).
    - isObsolete:** false
    - synonym:** none
  - obsolete\_cellular\_component - obsolete\_cellular\_component

Concepts **1-10** of 16

**I**: Is a **P**: Part of **G**: Correspondence **O**: Other relationship

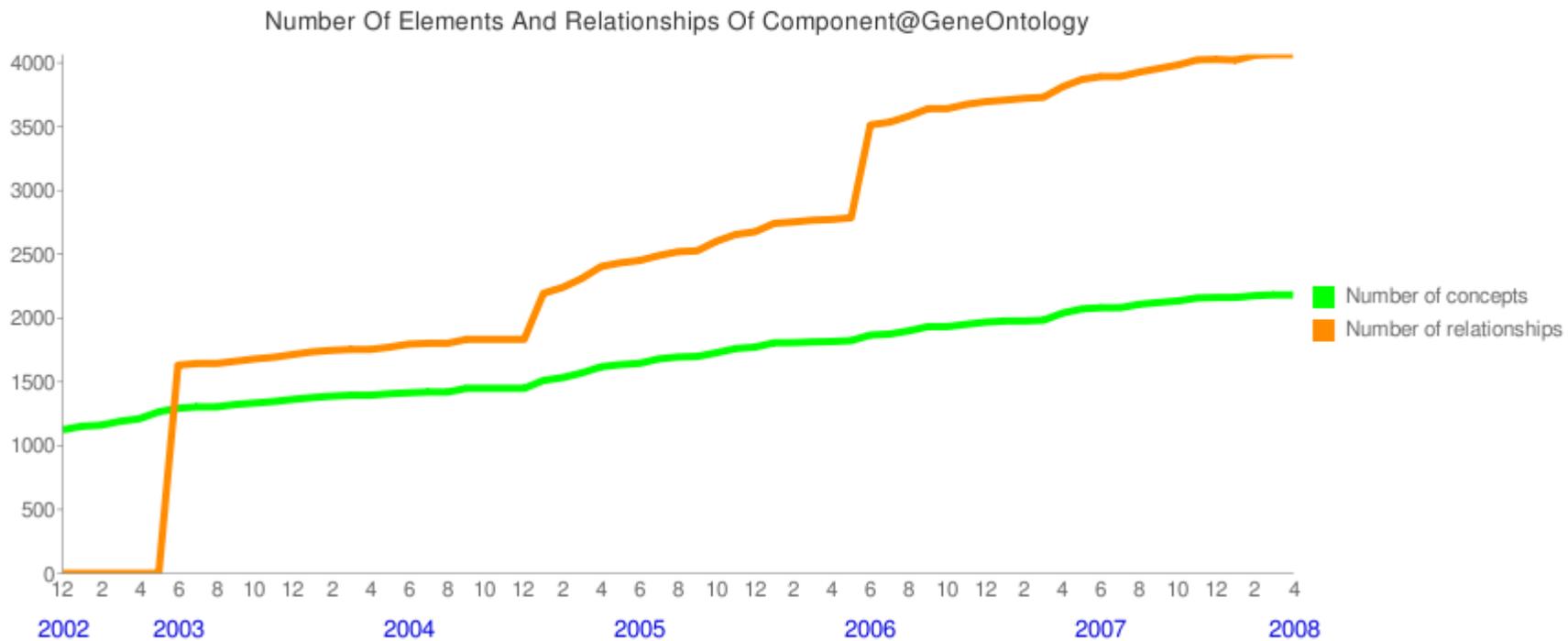
# 4. Evolutionsanalyse von Datenquellen



- **All-in Analyse**

- Liniendiagramm: Anzahl Elemente und Relationships aller Versionen

Source:

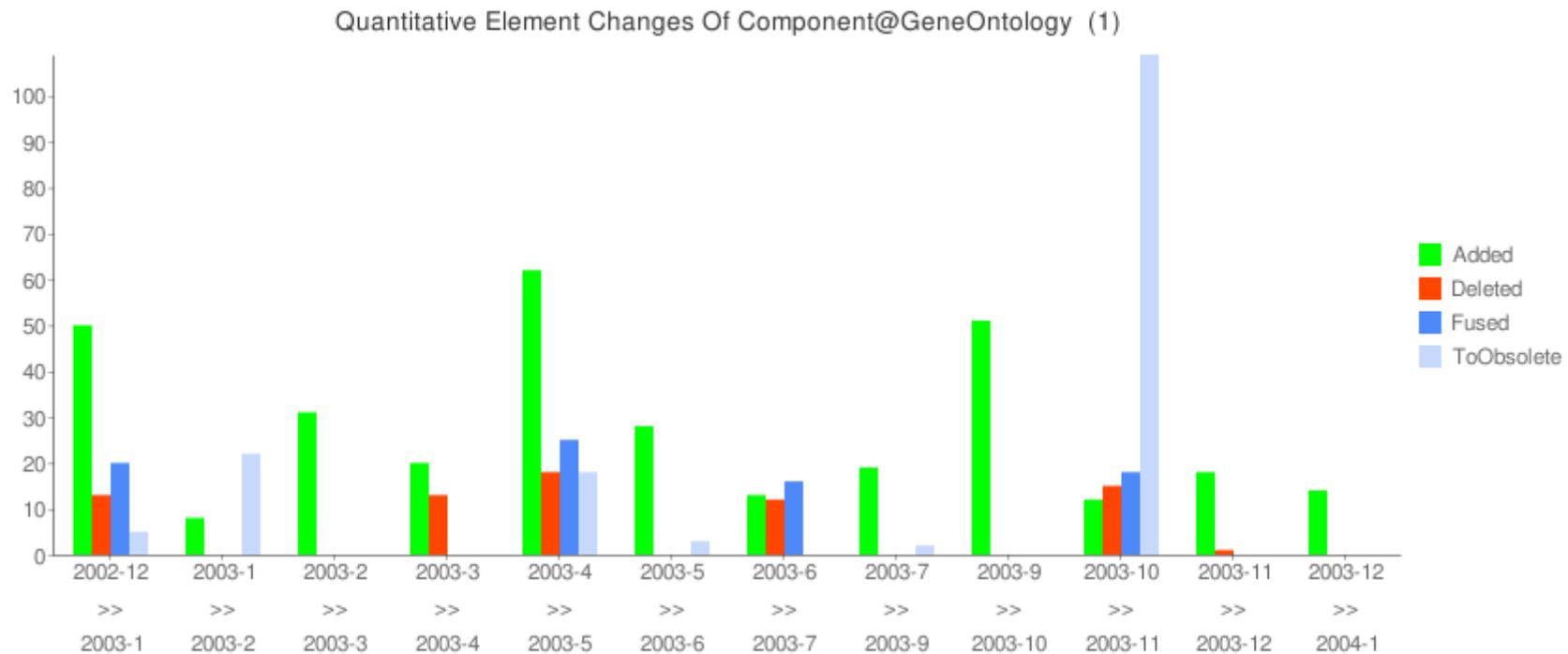


# 4. Evolutionsanalyse von Datenquellen



- **All-in Analyse**

- Balkengrafik: Anzahl add/delete/fuse/obsoletes zwischen Versionen



# 4. Evolutionsanalyse von Datenquellen



- All-in Analyse

Quantitative element changes of Component@GeneOntology					
From version	To version	Added concepts	Deleted concepts	Fused concepts	ToObsolete concepts
2002-12	2003-1	<u>29</u>	<u>1</u>	0	<u>5</u>
2003-1	2003-2	<u>8</u>	0	0	<u>22</u>
2003-2	2003-3	<u>31</u>	0	0	0
2003-3	2003-4	<u>20</u>	0	0	0
2003-4	2003-5	<u>62</u>	<u>8</u>	0	<u>18</u>
2003-5	2003-6	<u>28</u>	0	0	<u>3</u>
2003-6	2003-7				0
2003-7	2003-9				<u>2</u>
2003-9	2003-10				0
2003-10	2003-11				<u>109</u>
2003-11	2003-12				0
2003-12	2004-1				0
2004-1	2004-2				0
2004-2	2004-3				0
2004-3	2004-5				0
2004-5	2004-6				<u>3</u>
2004-6	2004-7				0

accession	name
GO:0000930	gamma-tubulin complex
GO:0030690	Noc1p-Noc2p complex
GO:0042735	protein body
GO:0000931	gamma-tubulin large complex
GO:0000923	equatorial mitotic organizing center
GO:0000811	GINS complex
GO:0030692	Noc4p-Nop14p complex
GO:0030689	Noc complex
GO:0000924	gamma-tubulin ring complex, centrosomal
GO:0000502	proteasome complex (sensu Eukarya)

Concepts **1-10** of 28



# 4. Evolutionsanalyse von Datenquellen

- Fokussierte Analyse

Source:  Period:

Start-Version:  End-Version:

December, 2002						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

1. Auswahl einer Datenquelle
2. Auswahl einer Periode
3. Auswahl von Start- und Endversion
4. Berechnung von Kennzahlen
5. Anzeige der Kennzahlenwerte und Werteverläufe  
(Grün für Unterschiede zu All-in Analyse)

Average additions per month	Average deletions per month	Average fusions per month	Average toObsolete per month
17.73	0.29	1.14	2.80

# 4. Explorative Analyse von Mappings



- **Keyword-Suche**

Domain Source:

Range Source:

Mapping:

Search in source:

Keyword:

1. Auswahl einer Domain- und Range Quelle
2. Auswahl eines verfügbaren Mappings
3. Auswahl einer Anfangsquelle
4. Eingabe von Suchwörtern
5. Darstellung des Mappings

# 4. Explorative Analyse von Mappings



- **Keyword-Suche**

🗄️ **Domain:** Function@GeneOntology - 2004-02

🗄️ ⓘ GO:0003674 - molecular\_function (0)

🗄️ ⓘ GO:0004871 - signal transducer activity (36) ↓

🗄️ ⓘ GO:0005057 - receptor signaling protein activity (16) ↓

🗄️ ⓘ GO:0005066 - transmembrane receptor protein tyrosine kinase signaling protein activity (1) ↓

🗄️ ⓘ **GO:0005068 - transmembrane receptor protein tyrosine kinase adaptor protein activity (3)** ↓

🗄️ ⓘ GO:0005887 - integral to plasma membrane

🗄️ ⓘ GO:0005886 - plasma membrane

🗄️ ⓘ GO:0005737 - cytoplasm

🗄️ ⓘ **GO:0005887 - integral to plasma membrane**

🗄️ **Range:** Component@GeneOntology - 2004-02

🗄️ ⓘ GO:0005575 - cellular\_component (0)

🗄️ ⓘ GO:0005623 - cell (5) ↓

🗄️ ⓘ **GO:0016020 - membrane (360)** ↓

🗄️ ⓘ **GO:0016021 - integral to membrane (568)** ↓

🗄️ ⓘ **GO:0005887 - integral to plasma membrane (424)** ↓

🗄️ ⓘ GO:0005886 - plasma membrane (206) ↓

🗄️ ⓘ **GO:0005887 - integral to plasma membrane (424)** ↓

ⓘ : Is a ⓘ : Part of ⓘ : Correspondence ⓘ : Other relationship

- **confidence:** 1.0
- **support:** 2

Show

# 4. Explorative Analyse von Mappings

- Mapping Navigation



Domain: Function@GeneOntology - 2004-02

- GO:0003674 - molecular\_function (0)
- GO:0005198 - structural molecule activity (56) ↓
- GO:0008369 - obsolete molecular function (0)
- GO:0004871 - signal transducer activity (36) ↓
- GO:0003754 - chaperone activity (36) ↓
- GO:0005554 - molecular\_function unknown (98) ↓
- GO:0005215 - transporter activity (58) ↓
- GO:0030188 - chaperone regulator activity (0)
- GO:0045182 - translation regulator activity (2) ↓
- GO:0045183 - translation factor activity, non-n...
- GO:0030371 - translation repressor activity (1)
- GO:0008135 - translation factor activity, nucleic acid binding (6) ↓
- GO:0005634 - nucleus**

**Correspondences**  
GO:0005634; GO:0005737

- **definition:** A membrane-bounded organelle of eukaryotic cells that contains the chromosomes. It is the primary site of DNA replication and RNA synthesis in the cell.
- **isObsolete:** false
- **synonym:** none

---

- **confidence:** 1.0
- **support:** 3

# 4. Evolutionsanalyse von Mappings



- Fokussierte Analyse

Domain Source:

Range Source:

1. Mapping:

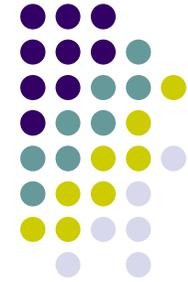
2. Mapping:

Mapping Name	Mapping Version	Number of correspondences	Coverage for domain	Coverage for range
GO_Function-GO_Components on Ensembl26(hsa)	2004-02 - 2004-02	8389	0.25	0.30
GO_Function-GO_Components on Ensembl25(hsa)	2004-02 - 2004-02	7279	0.23	0.28

From mapping	To mapping	Added correspondences	Deleted correspondences
GO_Function-GO_Components on Ensembl26(hsa) 2004-02 - 2004-02	GO_Function-GO_Components on Ensembl25(hsa) 2004-02 - 2004-02	<u>636</u>	<u>1746</u>

# 4. Mapping-Generierung

- Filterung



Domain Source:

Range Source:

Mapping:

Support >=   Confidence >=

Number of correspondences	Number of domain elements	Number of range elements	Coverage for domain	Coverage for range
1415	392	148	0.05	0.11

New mapping name:  File name:  .xml

File for download: [testfile.xml](#)

# 4. Mapping-Generierung

- Mengenmanipulation mit union, intersect and majority



Domain Source:

Range Source:

Mapping (multiselect with 'CTRL'):

- GO\_Function-GO\_Components on Ensembl25 (hsa): 2004-02 - 2004-02
- GO\_Function-GO\_Components on Ensembl26 (hsa): 2004-02 - 2004-02
- GO\_Function-GO\_Components on Ensembl27 (hsa): 2004-09 - 2004-09

Merge Method for Support:

Merge Method for Confidence:

Set Manipulation Method:

minNumber for majority (optional):

Number of correspondences	Number of domain elements	Number of range elements	Coverage for domain	Coverage for range
9025	1871	434	0.26	0.31

# 4. Mapping-Generierung

- Differenznermittlung



Domain Source:

Range Source:

1. Mapping:

2. Mapping:

Number of correspondences	Number of domain elements	Number of range elements	Coverage for domain	Coverage for range
636	278	152	0.04	0.11

# 4. Mapping-Generierung



- Komposition

- $O1 \leftrightarrow O2 \ \& \ O2 \leftrightarrow O3 \ \implies \ O1 \leftrightarrow O3$

Domain Mapping:   inverse

Range Mapping:   inverse

Number of correspondences	Number of domain elements	Number of range elements	Coverage for domain	Coverage for range
973403	1677	1677	0.23	0.23

# 4. Repository-Statistik



- Deskriptive Statistik zu integrierten Datenquellen

Source	First version	C  (first version)	Last version	C  (last version)	Number of versions	Average number of elements	Explorative Analysis
Concept@ChemicalEntitiesOfBiomedicalInterest	2004-10	10236	2008-5	19360	31	14,040.52	<input type="button" value="GO"/>
Component@GeneOntology	2002-12	1124	2008-4	2182	57	1,698.86	<input type="button" value="GO"/>
Function@GeneOntology	2002-12	5298	2008-4	8827	57	7,499.47	<input type="button" value="GO"/>
Process@GeneOntology	2002-12	6741	2008-4	15131	57	10,601.68	<input type="button" value="GO"/>

- Deskriptive Statistik zu Mappings

Mapping	Domain source	Range source	Number of correspondences	Number of domain elements	Number of range elements	Explorative analysis
GO_Function- GO_Components on Ensembl25(hsa)	Function@GeneOntology 2004-02	Component@GeneOntology 2004-02	7279	1677	386	<input type="button" value="GO"/>



# 5. Zusammenfassung

- GOMMA als System zur flexiblen Verwaltung von Ontology Mappings sowie deren Analyse
  - Explorative Analyse bzw. Evolutionsanalyse von Datenquellen und Mappings
  - Filterung bestehender und Ableitung neuer Mappings, z.B. durch Operationen wie  $\text{compose}(M1, M2)$ ,  $\text{diff}(M1, M2)$ ,  $\text{union}(M1, M2)$ ,  $\text{intersect}(M1, M2)$ , ...
  - Verschiedene UI Funktionen zur Analyse der Mappingdaten unter Nutzung von Web 2.0 Techniken



**Herzlichen Dank**