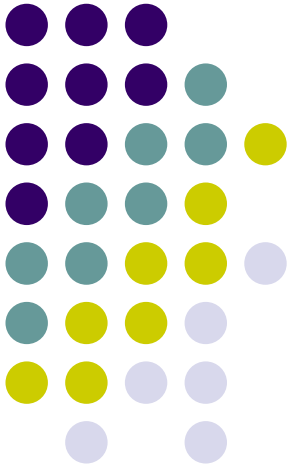
Diplomarbeit:

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

Bearbeiter: Shuangqing He
Betreuer: Toralf Kirsten, Michael Hartung



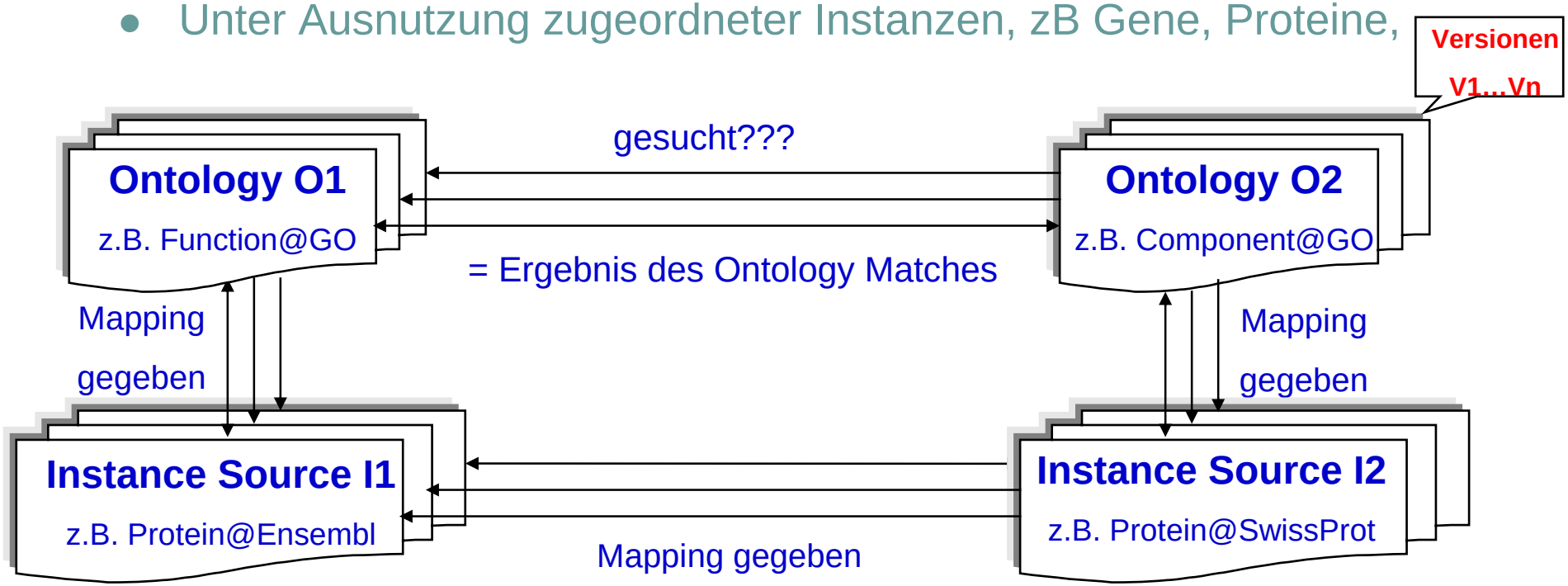
Universität Leipzig
Institut für Informatik
- Abteilung Datenbanken -

Agenda

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

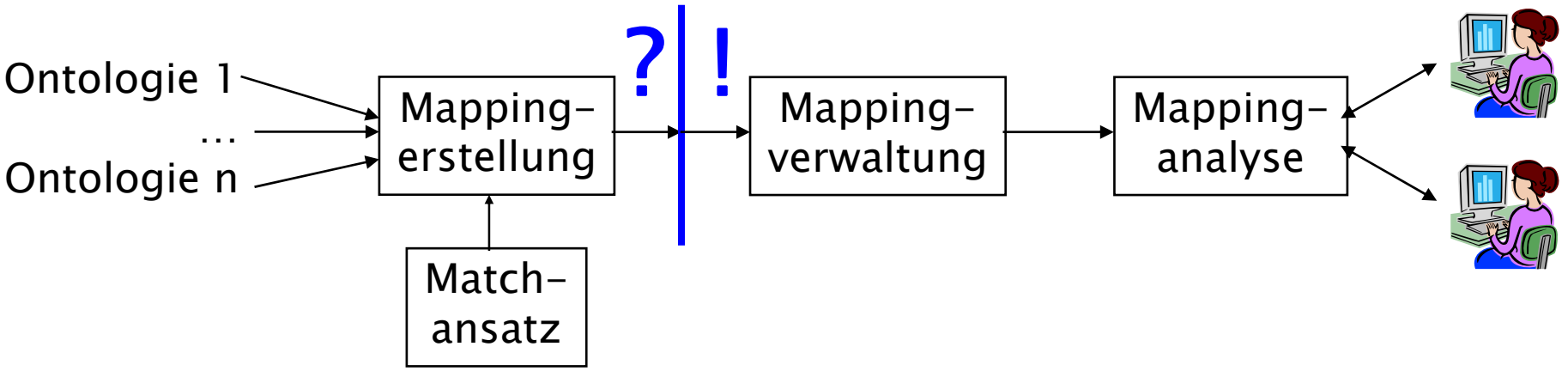
1. Motivation

- 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,



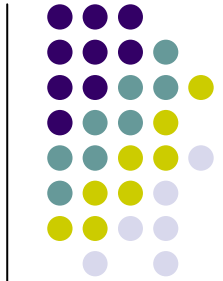
1. Motivation

- 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

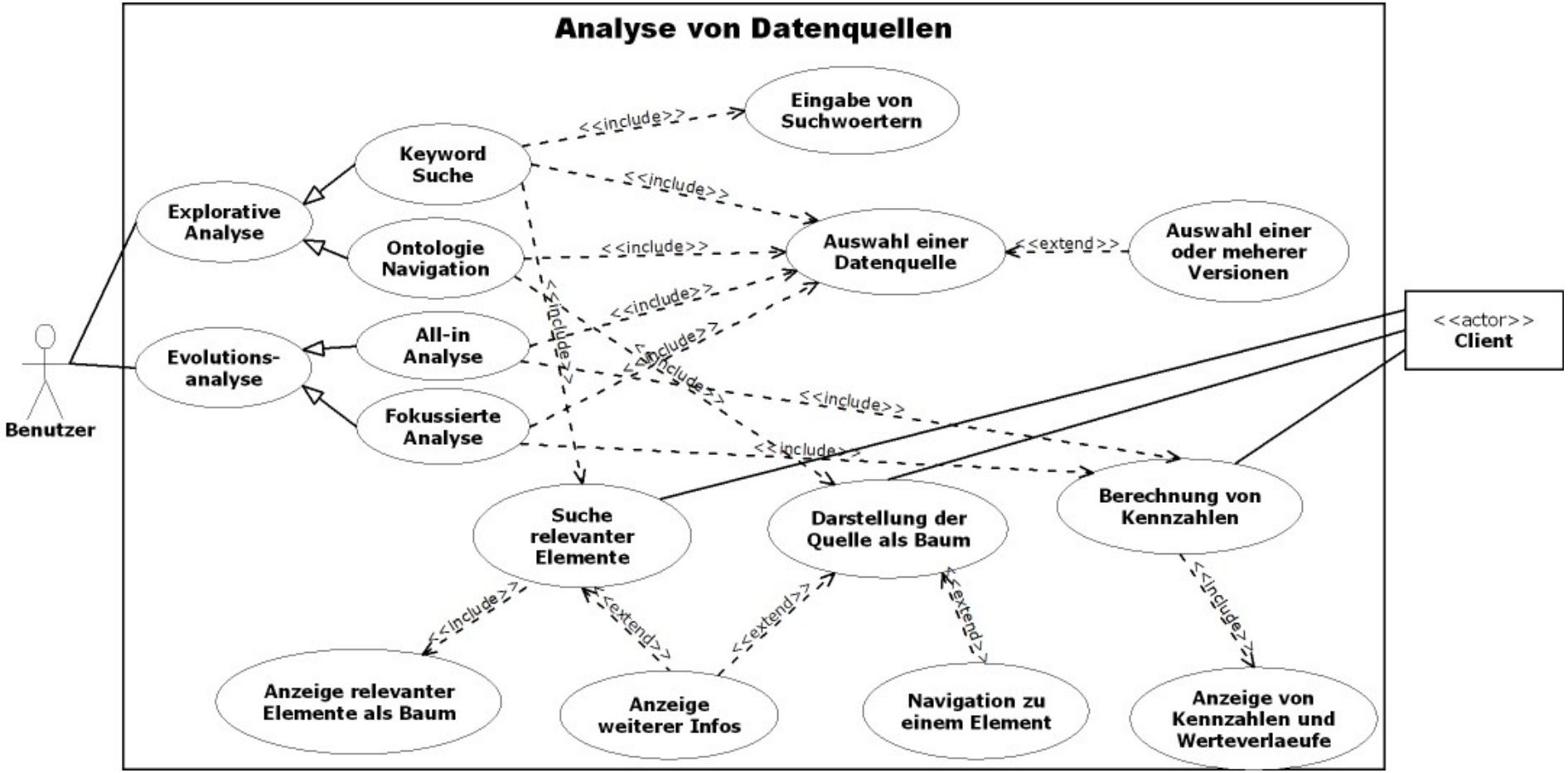


- Analyseanwendungen:
  - Explorative Analyse bzw. Evolutionsanalyse von Datenquellen
  - Explorative Analyse bzw. Evolutionsanalyse von Mappings
  - Mapping-Generierung
    - Filterung
    - Mengenmanipulation mit union, intersect und majority
    - Differenzenermittlung
    - Komposition

- Neuer Prototyp: GOMMA-Kernel, eine vorhandene API, die die Zugriffe auf das Mapping-Repository steuert/anbietet.

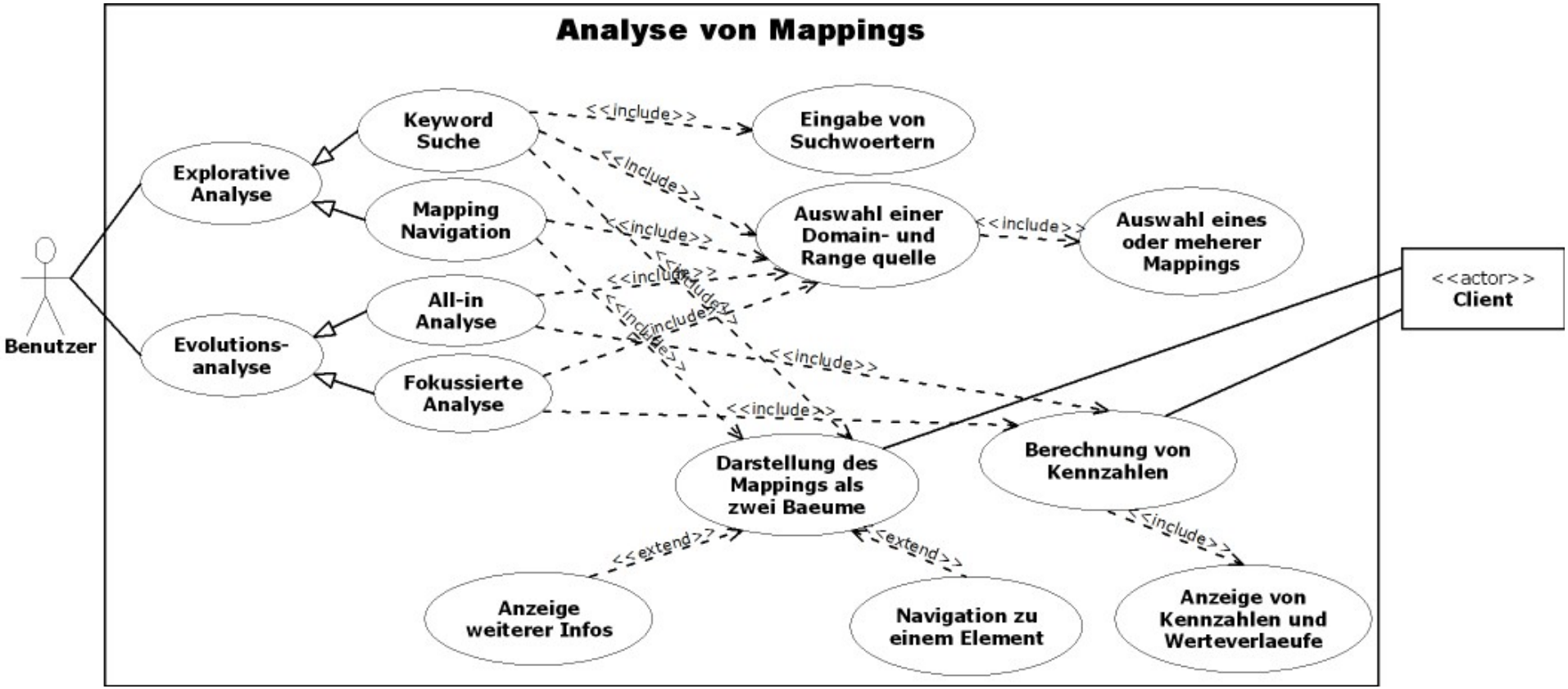
2. Nutzungsszenarien

- **Analyse von Datenquellen**



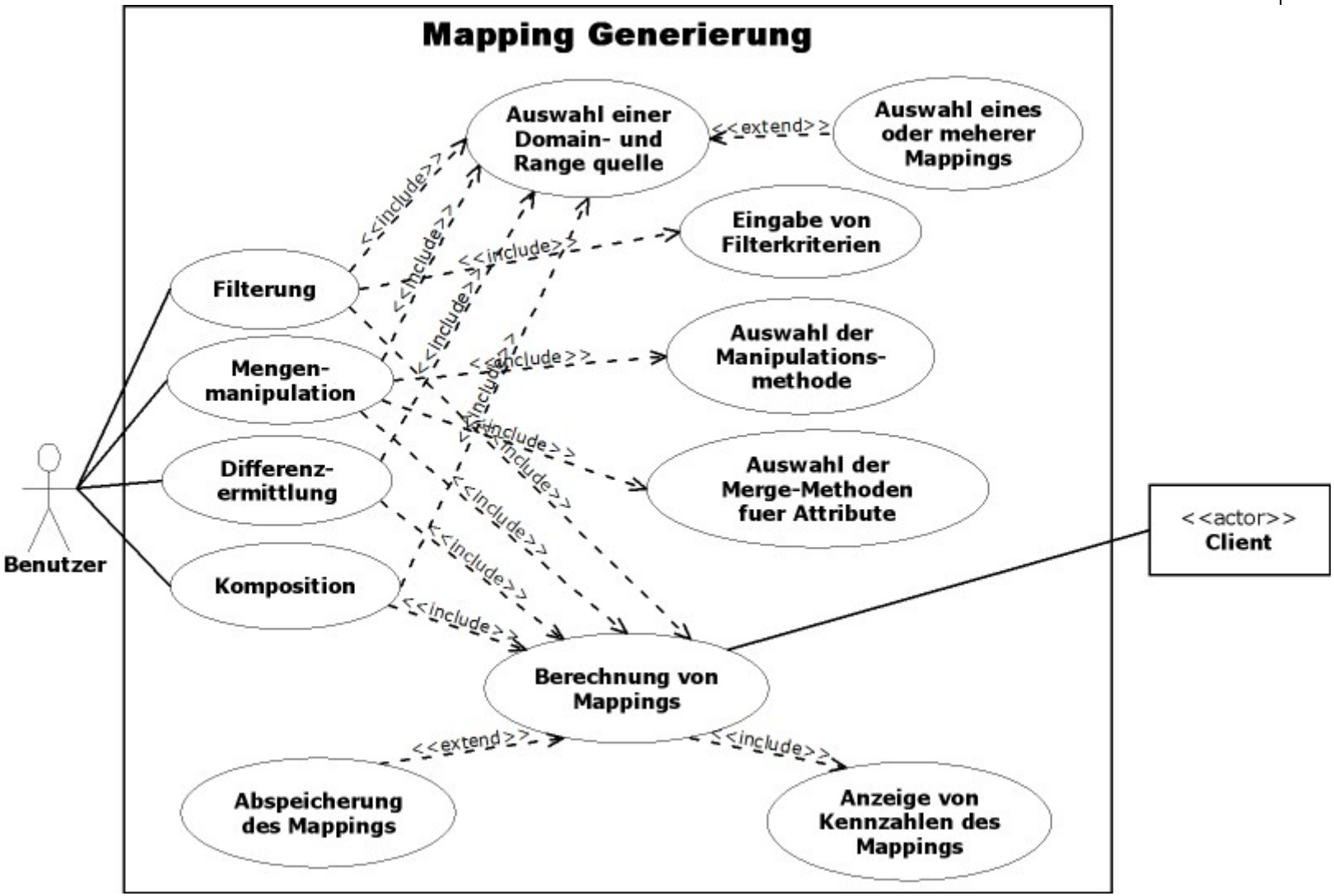
2. Nutzungsszenarien

- **Analyse von Mappings**



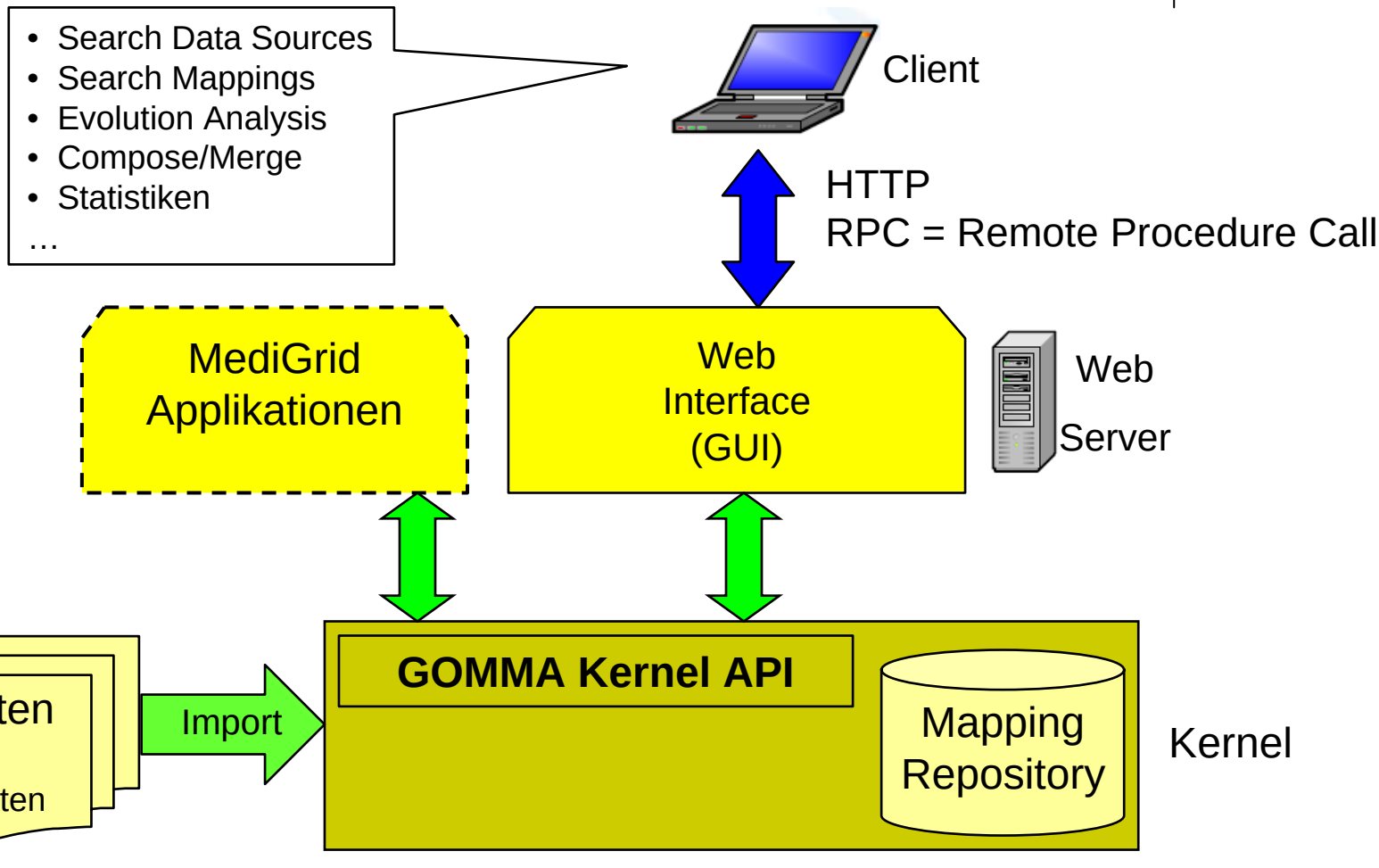
2. Nutzungsszenarien

- **Mapping Generierung**

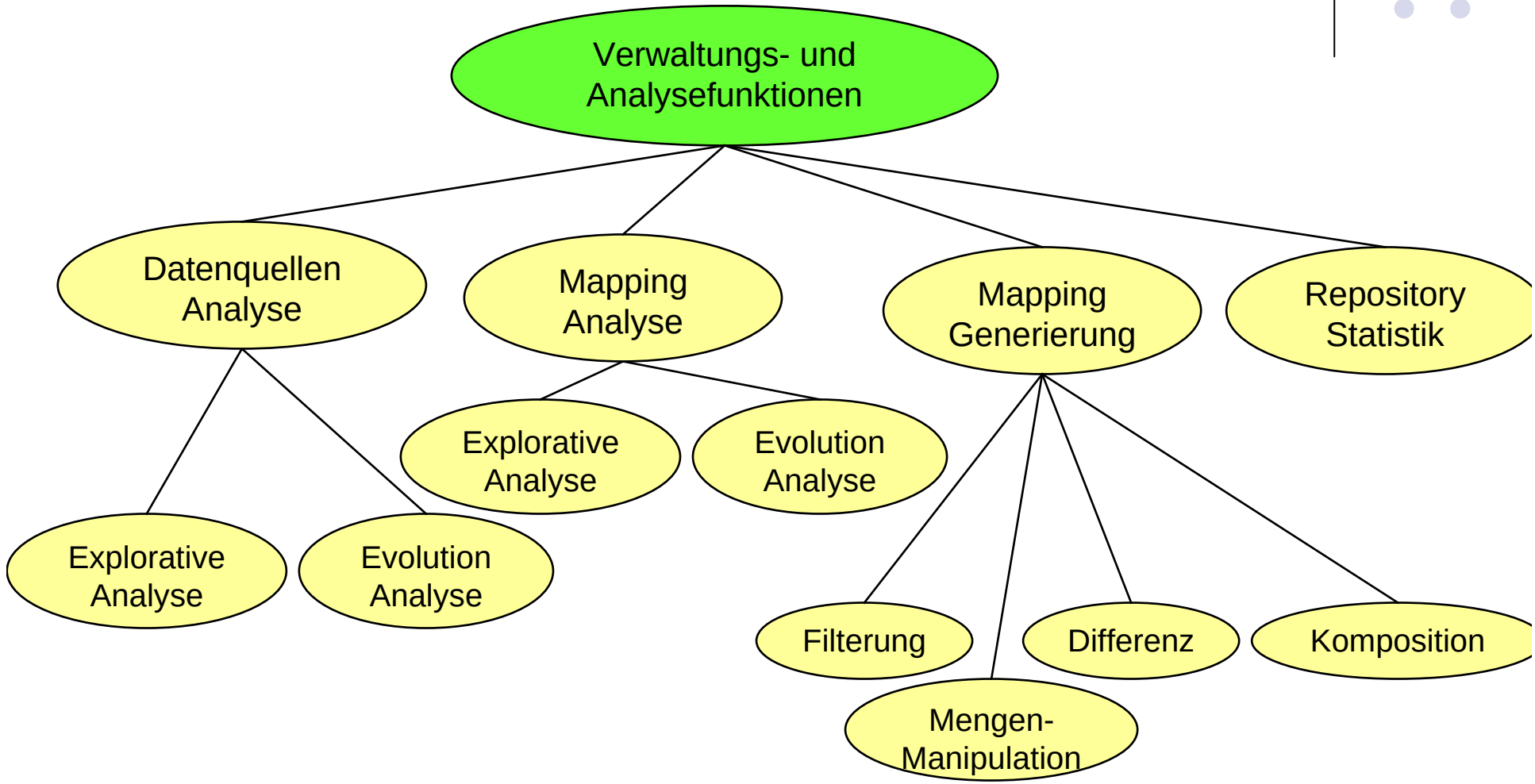


3. Systemarchitektur

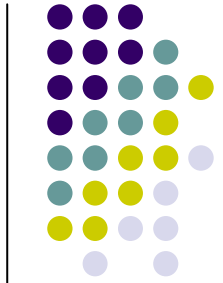
Komponenten des Generic Ontology Mapping Management & Analysis



3. Systemfunktionalitäten im Überblick

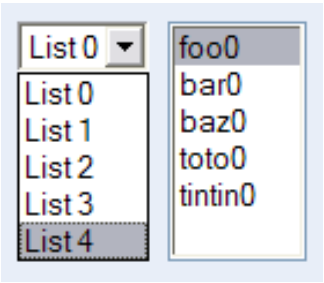


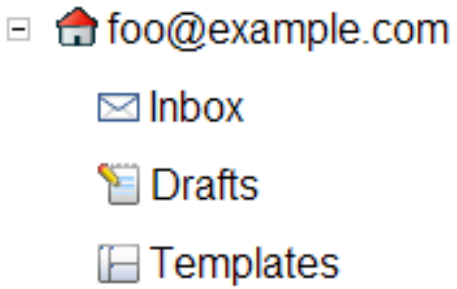
4. Implementierung: Überblick



- 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 ▸
Evolution Analysis ▸

Keyword Search
Ontology Navigation

The GOMMA provides mappings between different ontologies. Please use the menu on the top to explore the fu s underlying concepts:

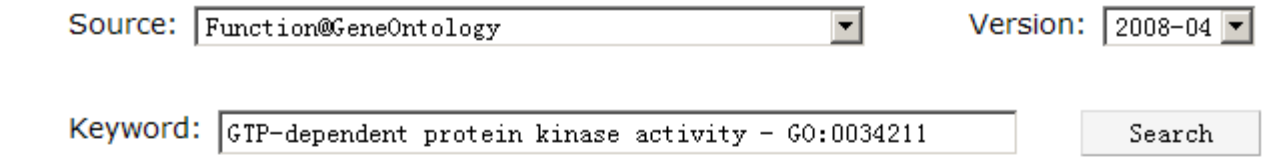
- **Home:** This page
- **Data Sources**: Explorative analysis and evolution analysis of data sources
- **Mappings**: Explorative analysis, evolution analysis and generation of mappings
- **Statistics**: Get statistics of available mappings and sources in GOMMA
- **Help**: Get help
- **Contact**: Feedback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

Current Version:

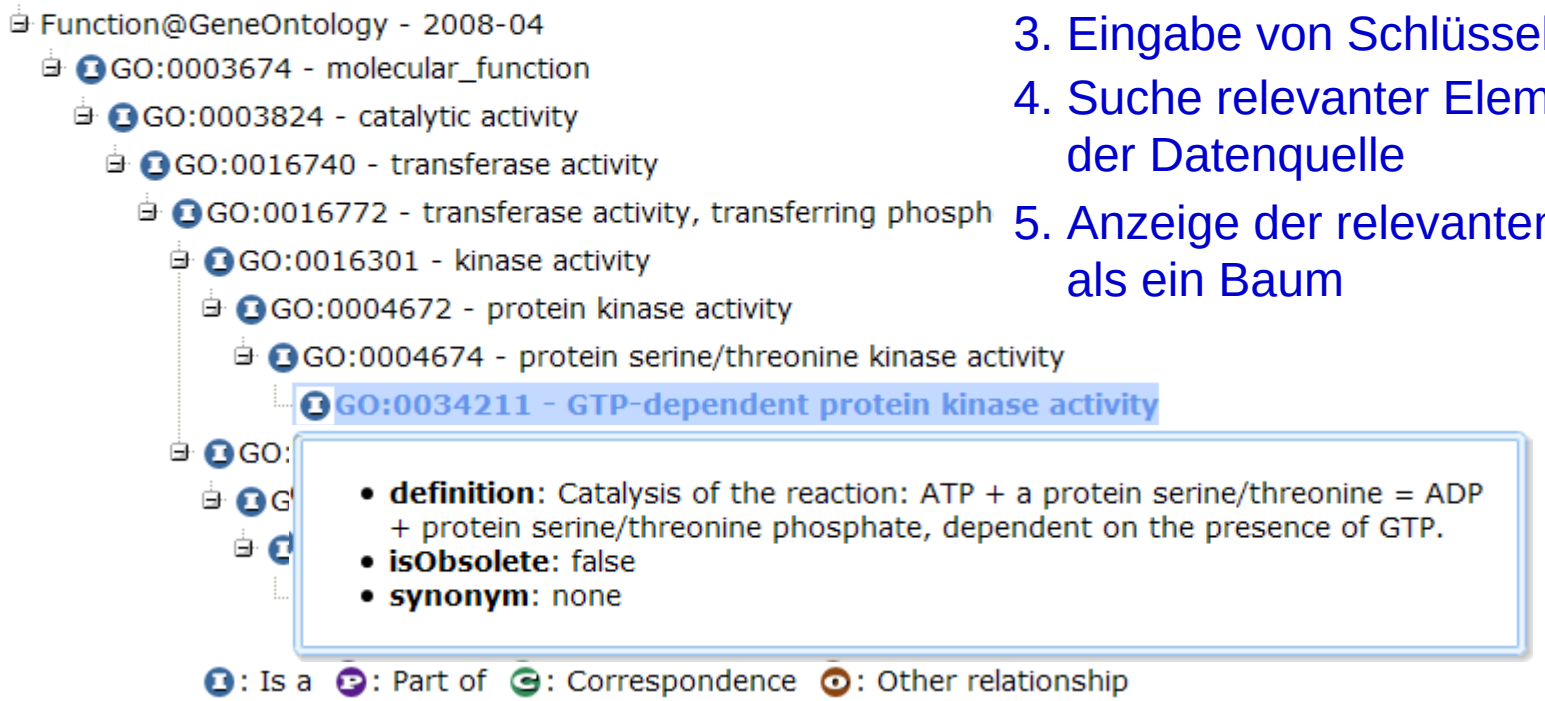
- Program version: 0.5
- Last update: **05.06.2008**

4. Explorative Analyse von Datenquellen

- **Keyword-Suche**

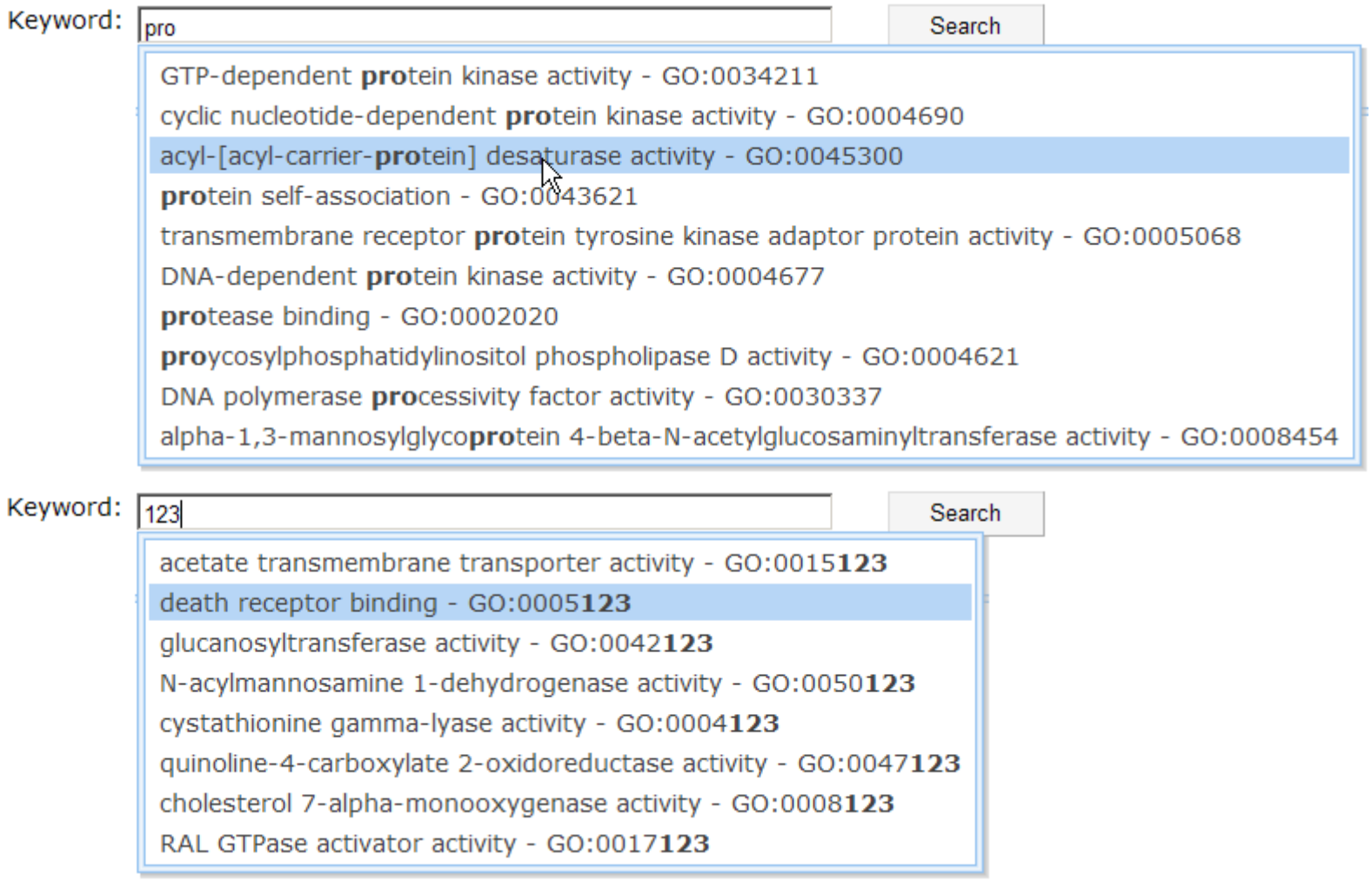


1. Auswahl einer Datenquelle
2. Auwahl einer Version
3. Eingabe von Schlüsselwörtern
4. Suche relevanter Elemente der Datenquelle
5. Anzeige der relevanten Elemente als ein Baum



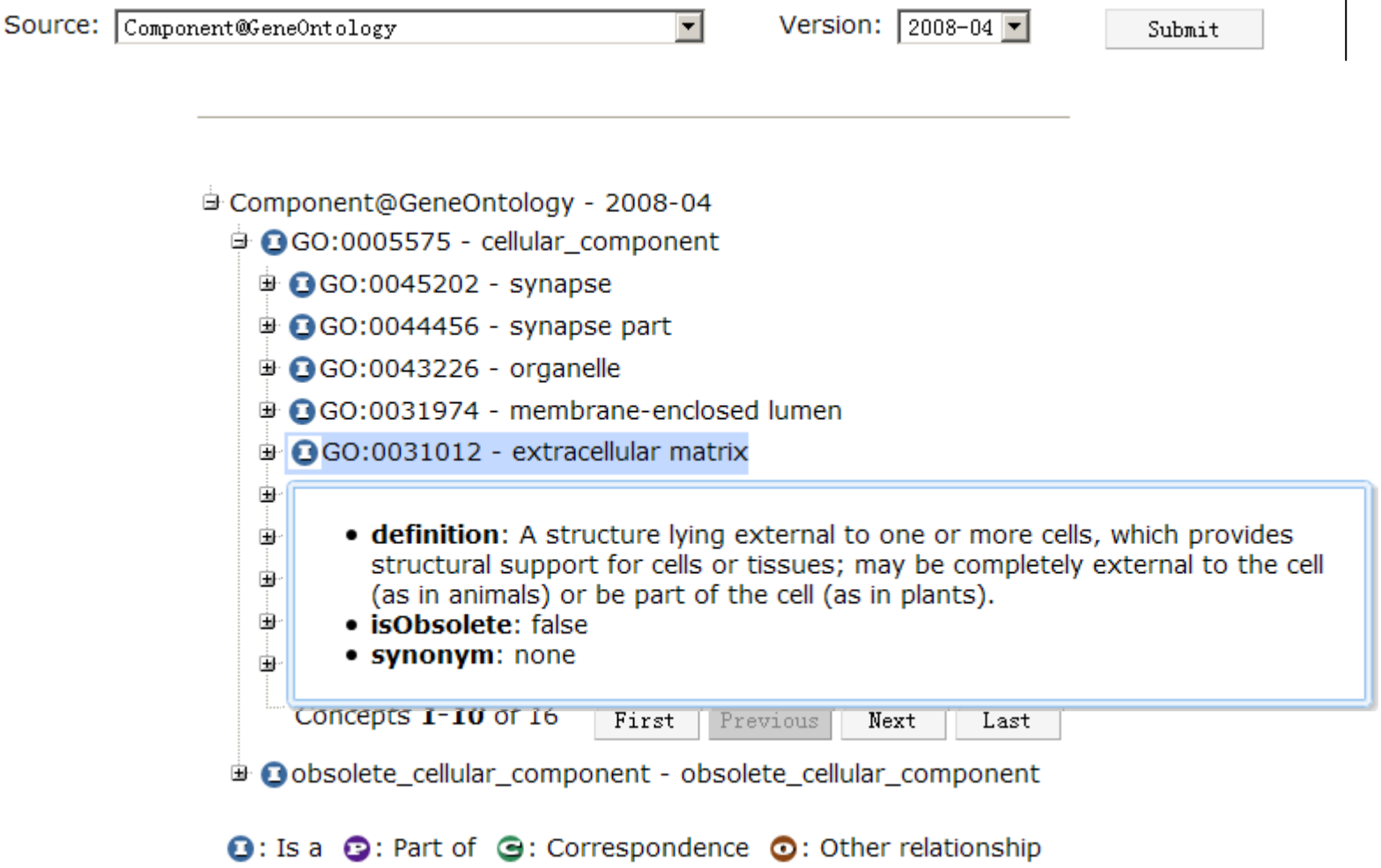
4. Anfrageformulierung

- **Anfrageformulierung (autocomplete):**



4. Explorative Analyse von Datenquellen

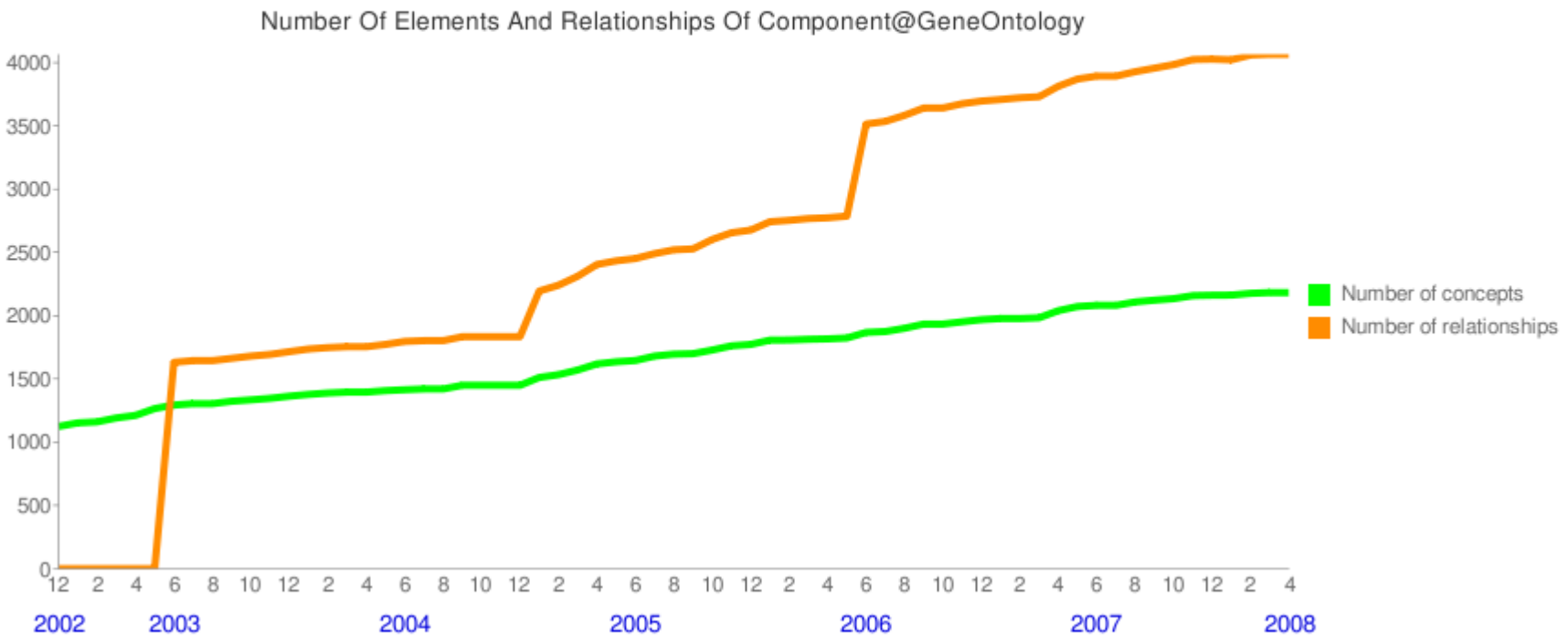
- **Ontologie Navigation**



4. Evolutionsanalyse von Datenquellen

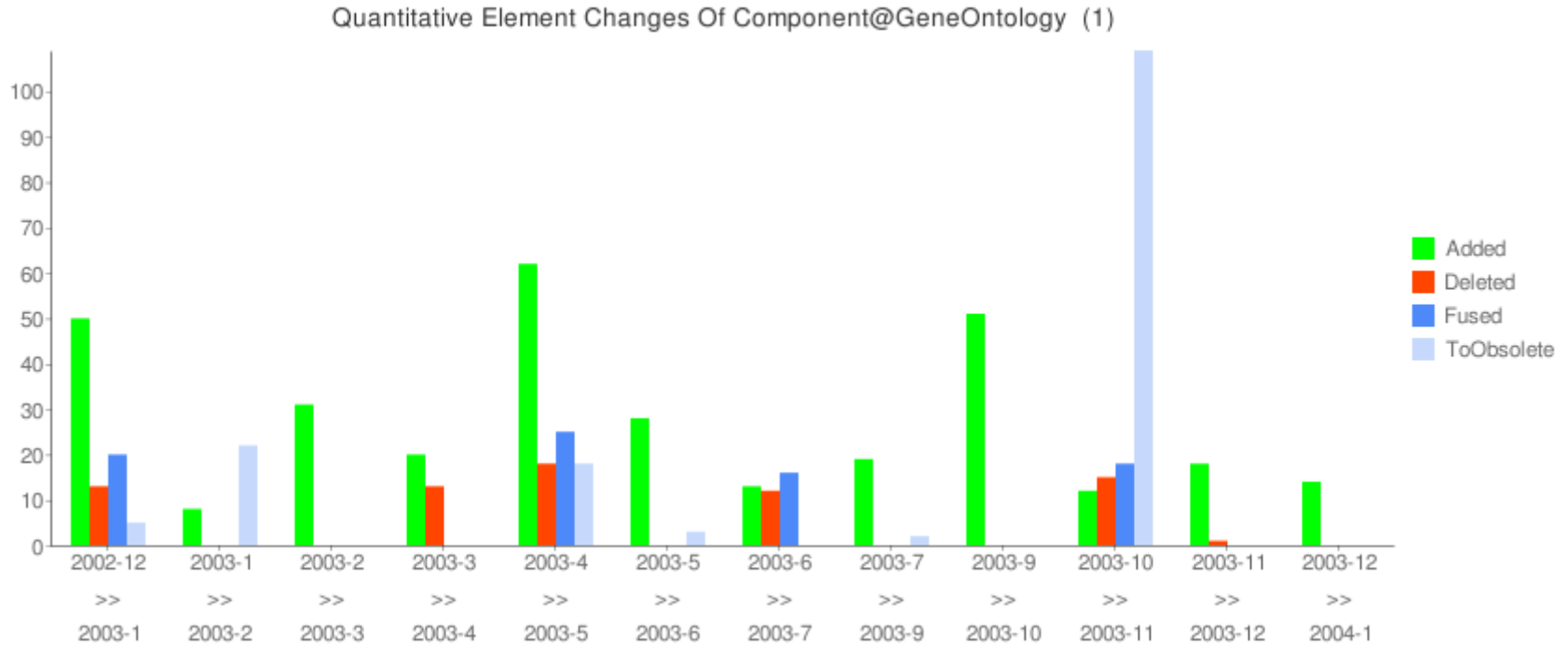
- **All-in Analyse**
  - Liniendiagramm: Anzahl Elemente und Relationships aller Versionen

Source: Component@GeneOntology Submit



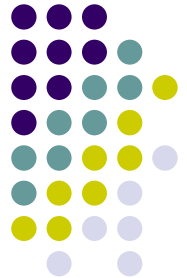
4. Evolutionsanalyse von Datenquellen

- **All-in Analyse**
  - Balkengrapfik: Anzahl add/delete/fuse/obsoletes zwischen Versionen



4. Evolutionsanalyse von Datenquellen

- **All-in Analyse**



Quantitative element changes of Component@GeneOntology

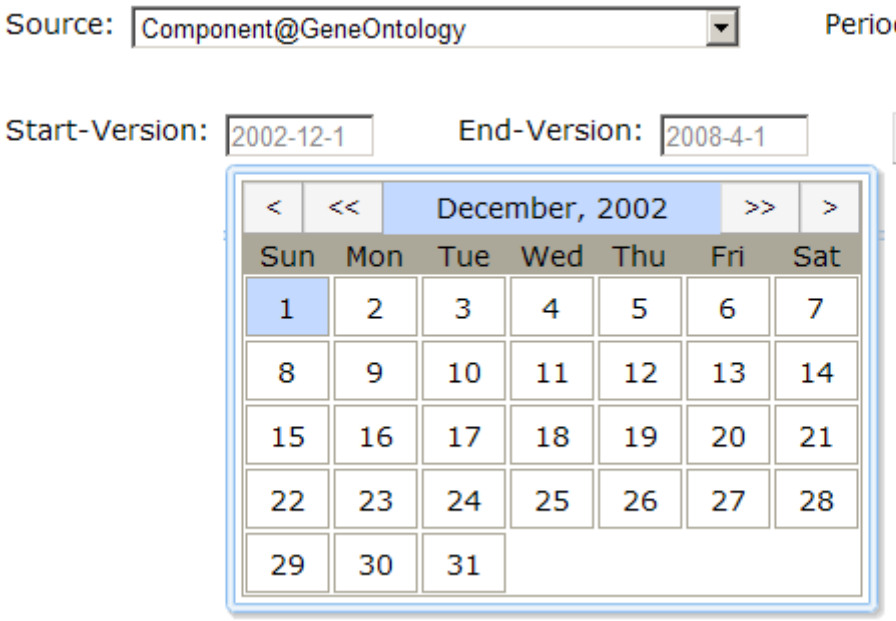
From version	To version	Added concepts	Deleted concepts	Fused concepts	ToObsolete concepts
2002-12	2003-1	29	1	0	5
2003-1	2003-2	8	0	0	22
2003-2	2003-3	31	0	0	0
2003-3	2003-4	20	0	0	0
2003-4	2003-5	62	8	0	18
2003-5	2003-6	28	0	0	3
2003-6	2003-7				0
2003-7	2003-9				2
2003-9	2003-10				0
2003-10	2003-11				109
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				3
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 First Previous Next Last

4. Evolutionsanalyse von Datenquellen

- **Fokussierte Analyse**



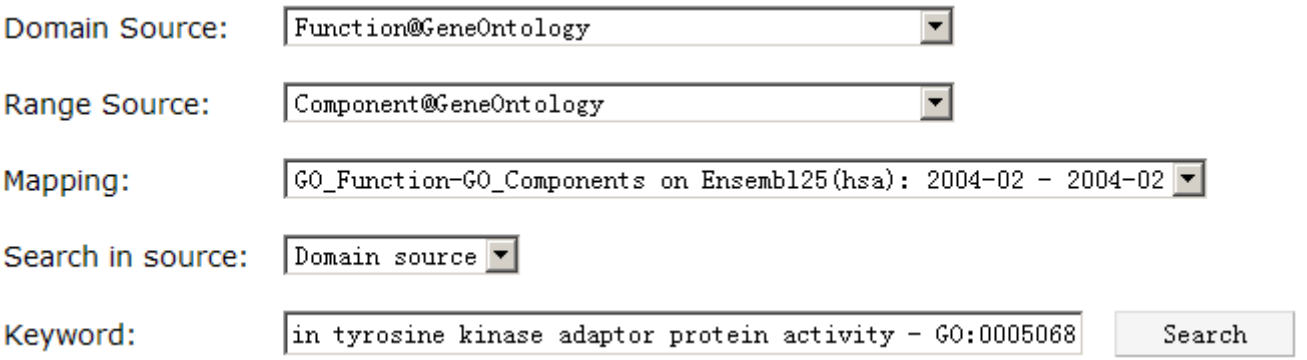
1. Auswahl einer Datenquelle
2. Auwahl einer Periode
3. Auswahl von Start- und Endversion
4. Berechnung von Kennzahlen
5. Anzeige der Kennzahlenwerte und Werteverläufe

(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. Evplorative Analyse von Mappings

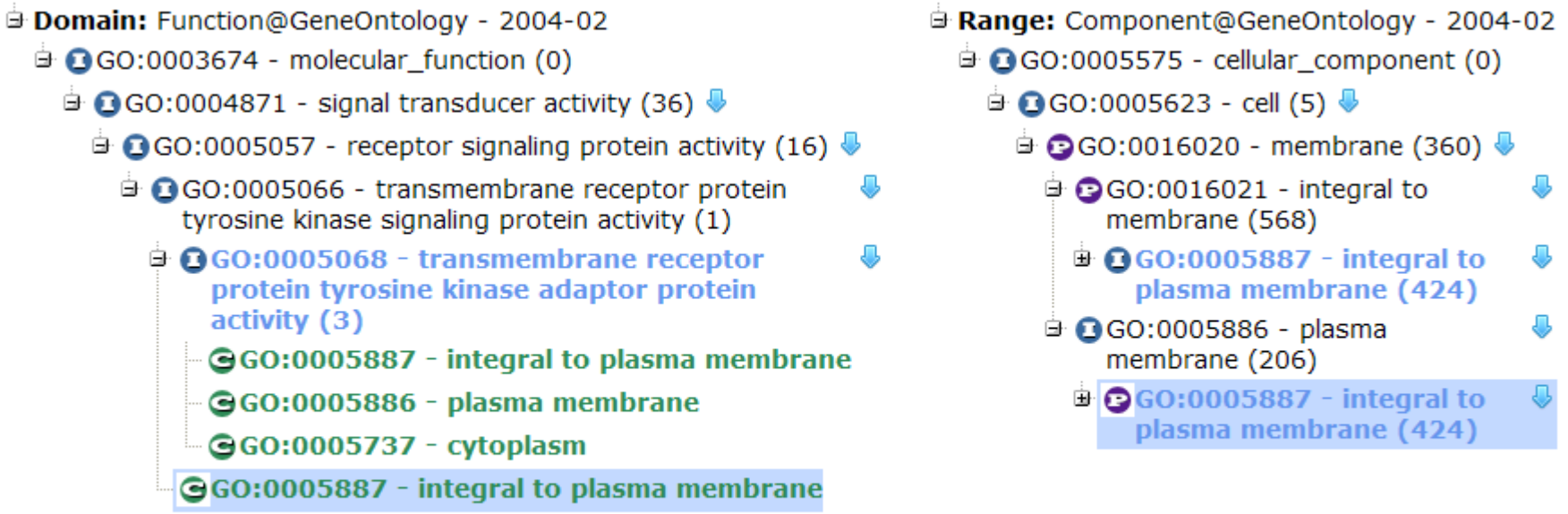
- **Keyword-Suche**



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

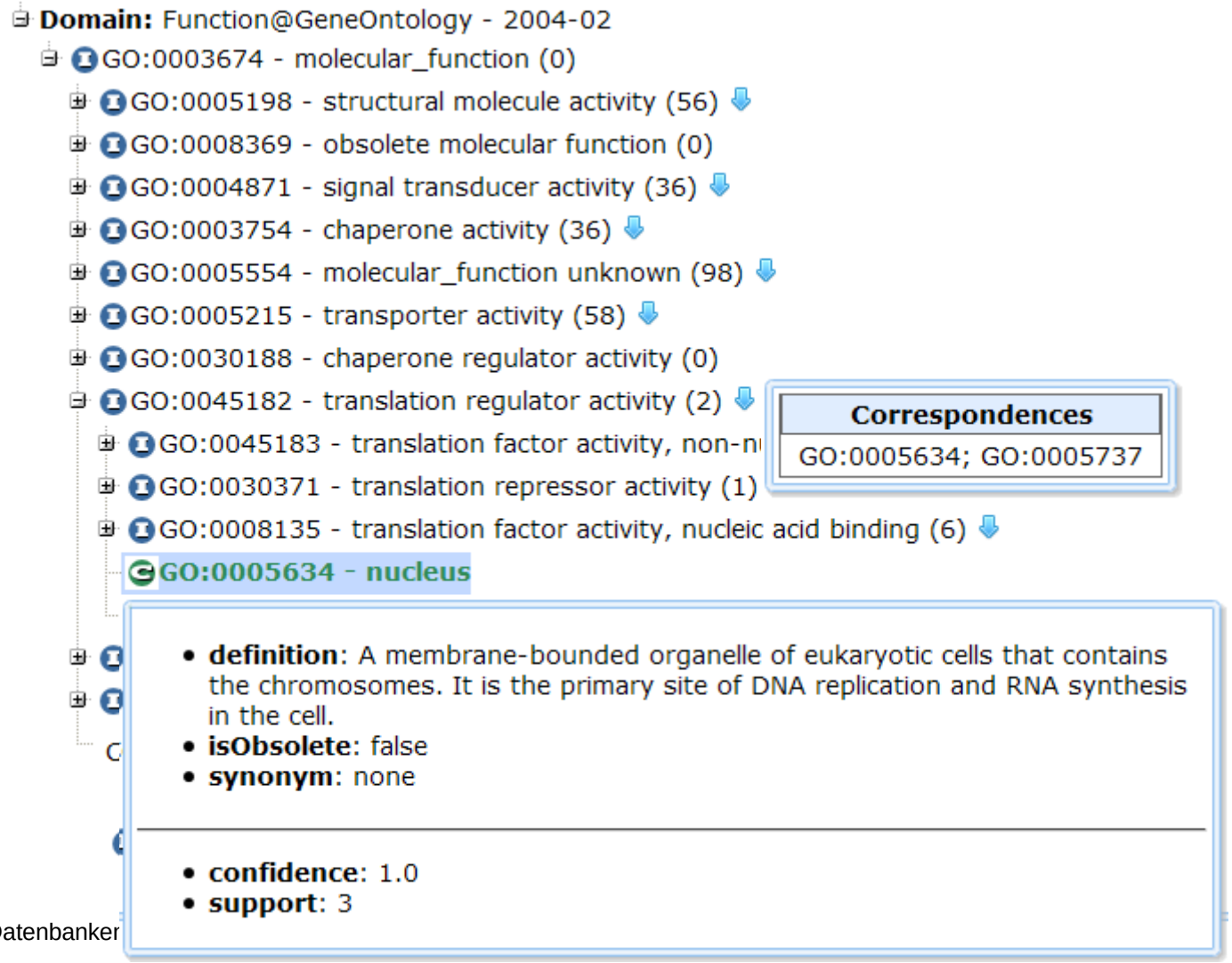
4. Evplorative Analyse von Mappings

- **Keyword-Suche**



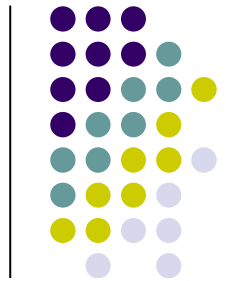
4. Evplorative Analyse von Mappings

- **Mapping Navigation**



4. Evolutionsanalyse von Mappings

- **Fokussierte Analyse**



Domain Source: Function@GeneOntology

Range Source: Component@GeneOntology

1. Mapping: GO_Function-GO_Components on Ensembl26(hsa): 2004-02 - 2004-02

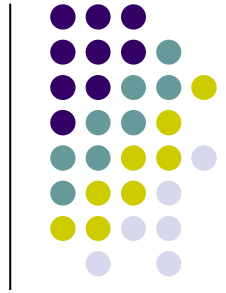
2. Mapping: GO_Function-GO_Components on Ensembl25(hsa): 2004-02 - 2004-02 Submit

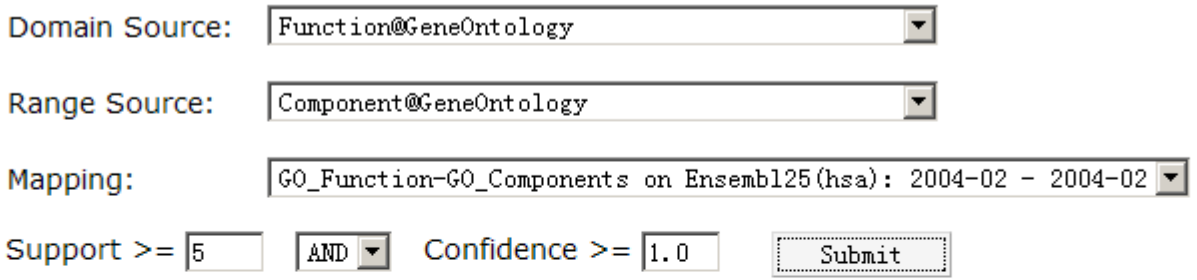
Mapping Name	Mapping Version	Number of correspondences	Coverage for domain	Coverage for range
GO_Function-GO_Components on Ensembl26(hsa)	2004-02 - 2004-02	8389	0.25	0.30
GO_Function-GO_Components on Ensembl25(hsa)	2004-02 - 2004-02	7279	0.23	0.28

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	636	1746

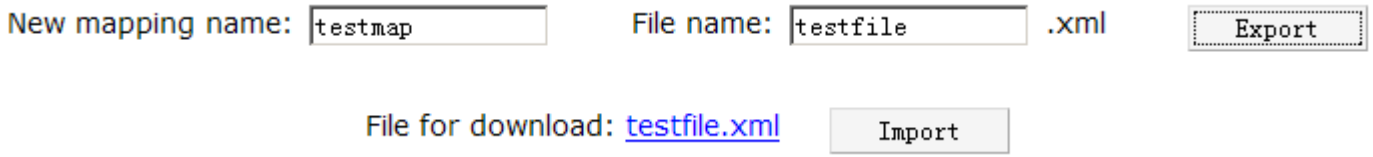
4. Mapping-Generierung

- Filterung



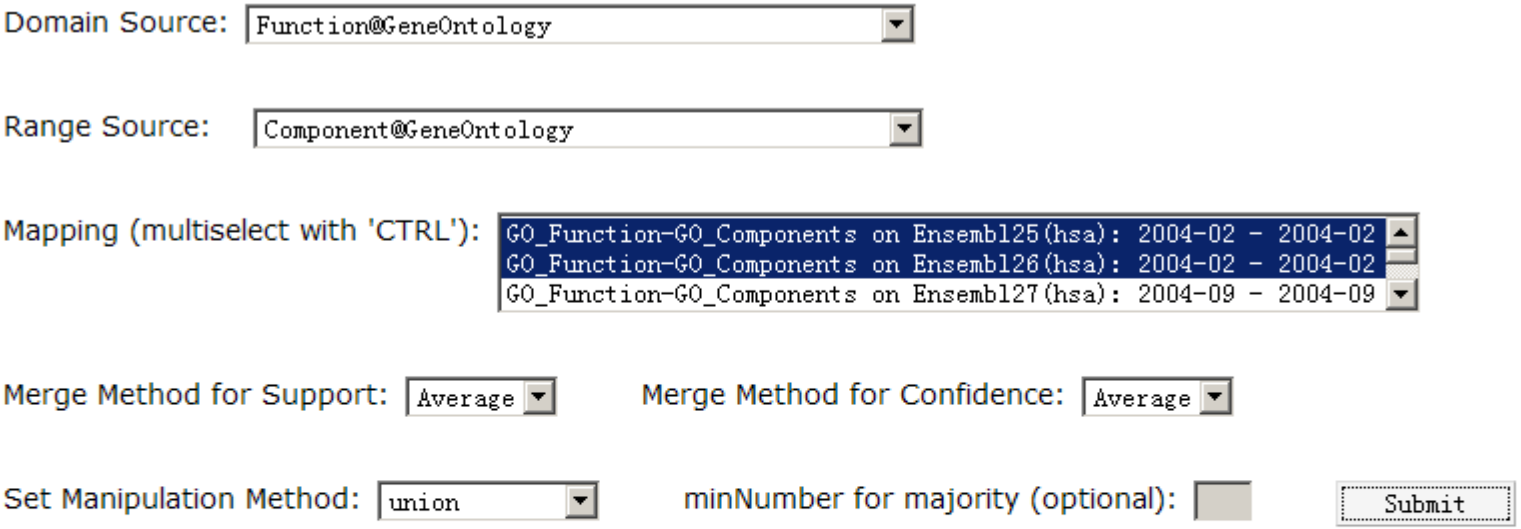


Number of correspondences	Number of domain elements	Number of range elements	Coverage for domain	Coverage for range
1415	392	148	0.05	0.11



4. Mapping-Generierung

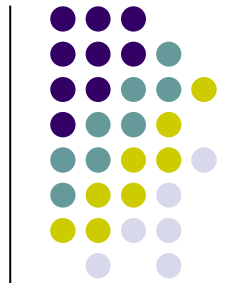
- Mengenmanipulation mit union, intersect and majority



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: Function@GeneOntology

Range Source: Component@GeneOntology

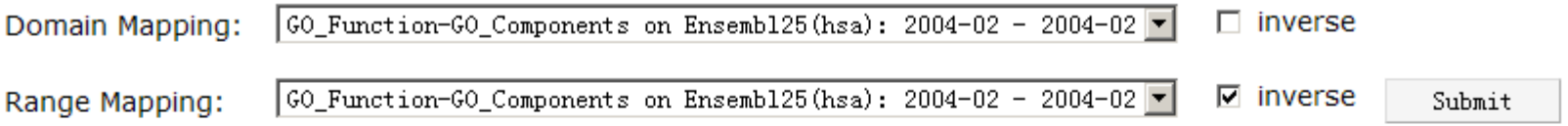
1. Mapping: GO_Function-GO_Components on Ensembl25(hsa): 2004-02 - 2004-02

2. Mapping: GO_Function-GO_Components on Ensembl26(hsa): 2004-02 - 2004-02 Submit

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

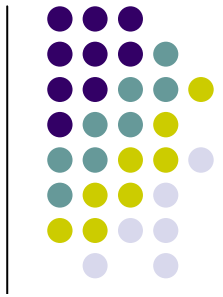
4. Mapping-Generierung

- Komposition
  - O1 ⟷ O2 & O2 ⟷ O3 ⟹ O1 ⟷ O3



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



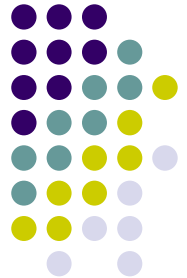
- Deskriptive Statistik zu integrierten Datenquellen

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	GO
Component@GeneOntology	2002-12	1124	2008-4	2182	57	1,698.86	GO
Function@GeneOntology	2002-12	5298	2008-4	8827	57	7,499.47	GO
Process@GeneOntology	2002-12	6741	2008-4	15131	57	10,601.68	GO

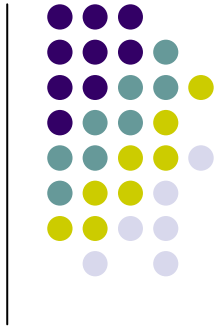
- Deskriptive Statistik zu Mappings

Mapping	Domain source	Range source	Number of correspondences	Number of domain elements	Number of range elements	Explorative analysis
GO_Function-GO_Component on Ensembl25(hsa)	Function@GeneOntology 2004-02	Component@GeneOntology 2004-02	7279	1677	386	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 compose(M1,M2), diff(M1,M2), union(M1,M2), intersect(M1,M2), …
  - Verschiedene UI Funktionen zur Analyse der Mappingdaten unter Nutzung von Web 2.0 Techniken



Herzlichen Dank