



A2-D4

Bioinformatics Demonstrators

Project title:	Reasoning on the Web with Rules and Semantics
Project acronym:	REWERSE
Project number:	IST-2004-506779
Project instrument:	EU FP6 Network of Excellence (NoE)
Project thematic priority:	Priority 2: Information Society Technologies (IST)
Document type:	D (deliverable)
Nature of document:	R (report)
Dissemination level:	PU (public)
Document number:	IST506779/Dresden/A2-D4/D/PU/b1
Responsible editors:	Gihan Dawelbait, Andreas Doms, Loic Royer
Reviewers:	Michael Schroeder
Contributing participants:	Dresden, Linköping
Contributing workpackages:	A2
Contractual date of deliverable:	28 Feb 2006
Actual submission date:	20 Feb 2006

Abstract

This deliverable presents demonstrators, which build upon the use cases specified in A2-D3 and which showcase bioinformatics applications using rules, reasoning and the web.

- GoProteins allows users to browse the GeneOntology and retrieve protein structure data from the web. It uses ontologies, deduction rules, reaction rules, databases and XML.
- Sambo integrates bio-ontologies such as the GeneOntology and Mesh. It deploys novel algorithms for concept mapping.
- BioRevise reasons over metabolic networks using techniques from belief revision.
- EarthFeed reads RSS feeds on the web, extracts locations and maps them onto a map of the world. The demonstrator uses XML and reactivity.
- GoPubMed allows users to explore PubMed search results with the GeneOntology.

The demonstrators can be shown at the annual review meeting.

Keyword List

Demonstrators, bioinformatics, pathways, PubMed, ontologies

Project co-funded by the European Commission and the Swiss Federal Office for Education and Science within the Sixth Framework Programme.

© REWERSE 2006.

Bioinformatics Demonstrators

Gihan Dawelbait^{Dre}, Andreas Doms^{Dre}, Patrick Lambrix^{Lin}, Loic Royer^{Dre},
Michael Schroeder^{Dre}

^{Dre} Technische Universität Dresden, Germany, ^{Lin} Linköpings universitet, Sweden

20 Feb 2006

Abstract

This deliverable presents demonstrators, which build upon the use cases specified in A2-D3 and which showcase bioinformatics applications using rules, reasoning and the web.

- GoProteins allows users to browse the GeneOntology and retrieve protein structure data from the web. It uses ontologies, deduction rules, reaction rules, databases and XML.
- Sambo integrates bio-ontologies such as the GeneOntology and Mesh. It deploys novel algorithms for concept mapping.
- BioRevise reasons over metabolic networks using techniques from belief revision.
- EarthFeed reads RSS feeds on the web, extracts locations and maps them onto a map of the world. The demonstrator uses XML and reactivity.
- GoPubMed allows users to explore PubMed search results with the GeneOntology.

The demonstrators can be shown at the annual review meeting.

Keyword List

Demonstrators, bioinformatics, pathways, PubMed, ontologies

Contents

1	Introduction	1
2	GoProtein	2
3	Sambo	6
4	BioRevise	9
5	EarthFeed and BioChemFeed	12
6	GoPubMed	14

1 Introduction

The deliverable A2-D4 sets out to deliver the following contents:

Demonstrators

With the requirement defined simple demonstrators will be produced. The demonstrators are mock-ups of future prototypes. They will demonstrate the main functionality of the specified applications and thus serve as a basis for the viability of the applications. Four applications will be pursued with the demonstrators:

- *Rule-based information integration and dissemination*
 - *Application: rule-based integration for structured data, e.g. protein interactions*
 - *Application: rule-based information dissemination for unstructured data, e.g. HIV information portal*
- *Rule-based querying for bioinformatics databases and tools*
 - *Application: rules, constraints, and ontologies to query structured data, e.g. constraints in sequence comparison*
 - *Application: rule and ontology-based search for unstructured data, e.g. biomedical literature*

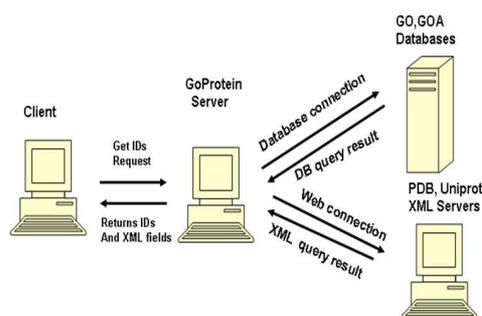
Here, we give a brief overview over the following demonstrators, which cover the applications sketched above.

- GoProteins allows users to browse the GeneOntology and retrieve protein structure data from the web. It uses ontologies, deduction rules, reaction rules, databases and XML.
- Sambo integrates bio-ontologies such as the GeneOntology and Mesh. It deploys novel algorithms for concept mapping.
- BioRevise reasons over metabolic networks using techniques from belief revision.
- EarthFeed reads RSS feeds on the web, extracts locations and maps them onto a map of the world. The demonstrator uses XML and reactivity.
- GoPubMed uses the GeneOntology to search the biomedical literature.

For GoProteins and Sambo the appendix contains details.

2 GoProtein

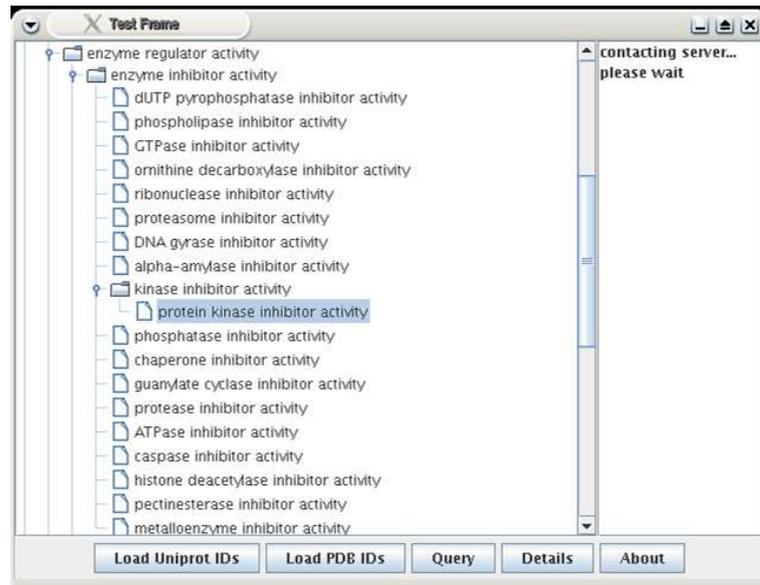
- **Name of the system:** GoProtein
- **Brief description:** The GoProtein tool browses the Gene Ontology (GO) and uses the Gene Ontology Annotation database(GOA) to retrieve relevant IDs for a selected term from the Protein Database Bank(PDB) and Uniprot protein databases. To retrieve more information from PDB and Uniprot, GoProtein connects to their respective URL, parses their XML entries and retrieves preselected fields. GoProtein implements the following workflow:



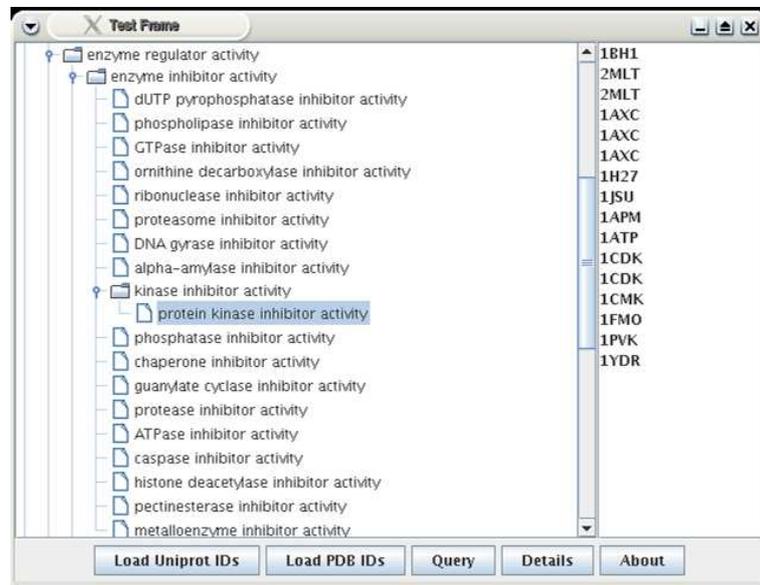
The user uses the client machine to issue a request to the server. The server then connects to the GO and GOA databases and process the query and returns the results to be displayed to the user. The user can request more information about the term by sending a query to PDB and Uniprot respective URLs and retrieve the query results.

- **Requirements:** GoProtein is written in Prova. It requires online internet access to PDB and MySQL access to the GO and GOA databases.
- **Example:** The following screenshots summarise how to show all the PDB entries associated with the GO term 'protein kinase inhibitor'.

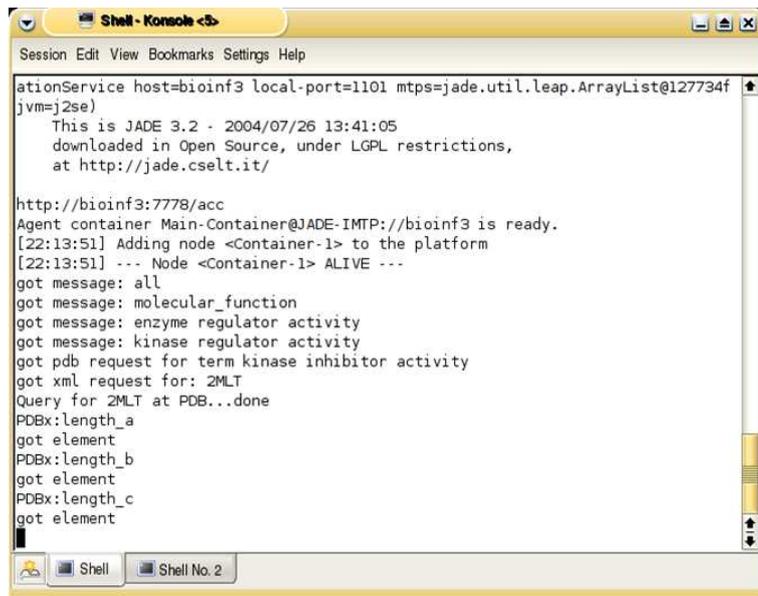
1. Select a term and wait for connection to database



2. Retrieval of associated PDB IDs for the selected term



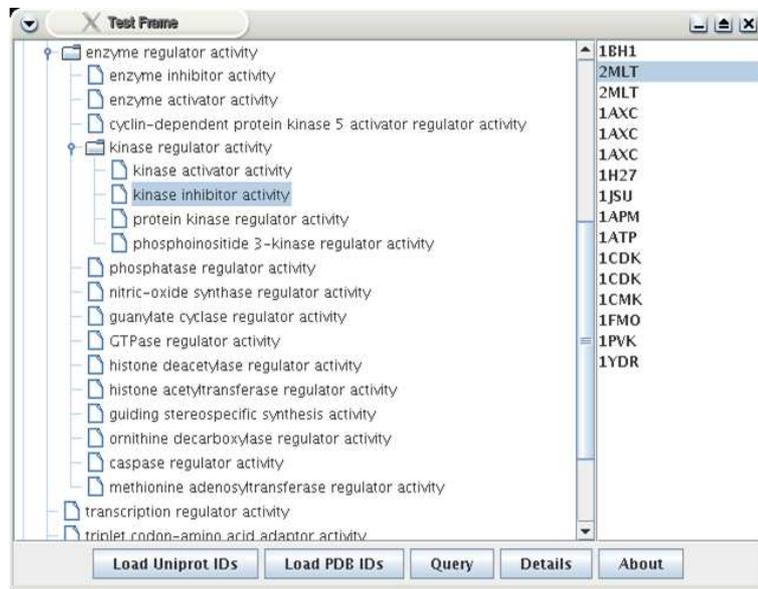
3. Message exchange between the two agents



```
Session Edit View Bookmarks Settings Help
ationService host=bioinf3 local-port=1101 mtps=jade.util.leap.ArrayList@127734f
jvm=j2se)
  This is JADE 3.2 - 2004/07/26 13:41:05
  downloaded in Open Source, under LGPL restrictions,
  at http://jade.cselt.it/

http://bioinf3:7778/acc
Agent container Main-Container@JADE-IMTP://bioinf3 is ready.
[22:13:51] Adding node <Container-1> to the platform
[22:13:51] --- Node <Container-1> ALIVE ---
got message: all
got message: molecular_function
got message: enzyme regulator activity
got message: kinase regulator activity
got pdb request for term kinase inhibitor activity
got xml request for: 2MLT
Query for 2MLT at PDB...done
PDBx:length_a
got element
PDBx:length_b
got element
PDBx:length_c
got element
```

4. Connecting to the web and processing XML pages to retrieve additional information about the term



Biological Activity	PDB IDs
enzyme regulator activity	1BHI
enzyme inhibitor activity	2MLT
enzyme activator activity	2MLT
cyclin-dependent protein kinase 5 activator regulator activity	1AXC
kinase regulator activity	1AXC
kinase activator activity	1AXC
kinase inhibitor activity	1H27
protein kinase regulator activity	1JSU
phosphoinositide 3-kinase regulator activity	1APM
phosphatase regulator activity	1ATP
nitric-oxide synthase regulator activity	1CDK
guanylate cyclase regulator activity	1CDK
CTPase regulator activity	1CMK
histone deacetylase regulator activity	1FMO
histone acetyltransferase regulator activity	1PVK
guiding stereospecific synthesis activity	1YDR
ornithine decarboxylase regulator activity	
caspase regulator activity	
methionine adenosyltransferase regulator activity	
transcription regulator activity	
triplelet codon-amino acid adaptor activity	

Buttons: Load Uniprot IDs, Load PDB IDs, Query, Details, About

5. Display of additional pages (sequence length)



• **Brief description how rules and other technologies are used:**

GoProtein is written in Prova, which integrates Java with deduction and reaction rules. GoProteins uses these rules as follows:

- Reaction rules specify the behaviour of the client, which displays the GUI, and the server agent, which handles database access and external access to the PDB server. The reaction rules include messaging in the agent communication language JADE.
- Deduction rules are used by the client agent to traverse the ontology.
- Database access is used by the server agent to access the GO database server.
- XML documents, which provide details of the protein structures, are processed with rules on the server side.
- The GUI is written entirely in Java, which is accessible from the Prova rule framework.

• **Brief summary of benefits:**

- Distributed application using XML, databases, reaction rules with messaging, and deduction rules for ontologies;
- Easy browsing of the Gene ontology tree;
- Instant access to informations from databases and web resources about protein annotation;
- Resilience to change (databases and web sites change often);
- Use of open source technologies;
- Location transparency (local, remote, mirrors);

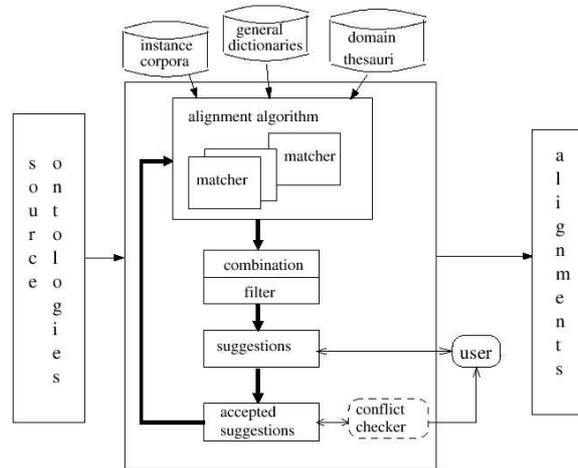


Figure 1: A general alignment strategy.

3 Sambo

- **Name of the system:** Sambo
- **Brief description:**

SAMBO is an ontology alignment and merging tool.

Figure 1 shows a general strategy for aligning two ontologies. An alignment algorithm receives as input two source ontologies. The algorithm can include several matchers. The matchers can implement strategies based on

- linguistic matching (e.g. string matching),
- structure-based strategies (e.g. similarity of children and parents),
- constraint-based approaches (e.g. inferring similarity from similar ranges of concepts),
- instance-based strategies (e.g. inferring similarity of GeneOntology terms from similarity of annotated proteins),
- strategies that use auxiliary information (e.g. domain knowledge from additional databases),
- or a combination of these.

The matchers calculate similarities between the terms from the different source ontologies. Alignment suggestions are then determined by combining and filtering the results generated by one or more matchers. The pairs of terms with a similarity value above a certain threshold are retained as alignment suggestions. By using different matchers and combining them and filtering the results in different ways we obtain different alignment strategies. The suggestions are then presented to the user who accepts or rejects them.

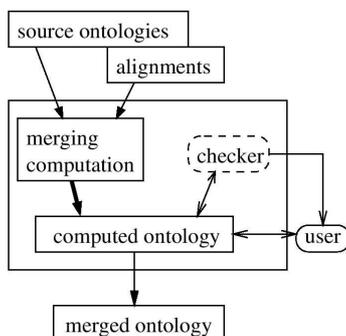


Figure 2: A general merging algorithm.

The acceptance and rejection of a suggestion may influence further suggestions. Further, a conflict checker is used to avoid conflicts introduced by the alignment relationships. The output of the alignment algorithm is a set of alignment relationships between terms from the source ontologies.

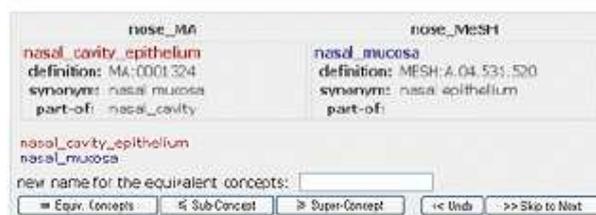
Figure 2 shows a simple merging algorithm. A new ontology is computed from the source ontologies and their identified alignment. The checker is used to avoid conflicts as well as to detect unsatisfiable concepts and, if so desired by the user, to remove redundancy.

- **Requirements:** Sambo is implemented in Java and deployed in Tomcat.
- **Example:** The following screenshots summarise a session in Sambo.

1. Choosing different matchers and assigning weights



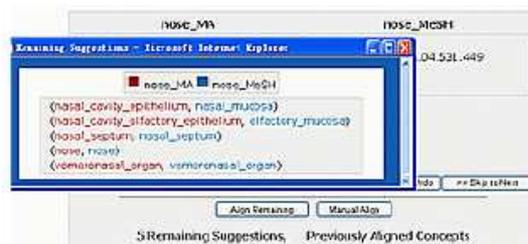
2. Alignment suggestion including definition/identifier, synonyms, relations



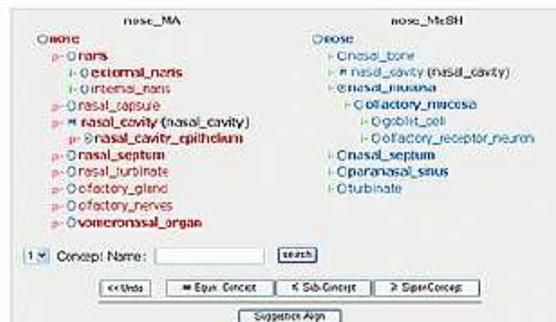
3. Rejecting a suggestion



4. Overview over remaining alignments



5. Manual alignment



- **Brief description how rules and other technologies are used:**

OWL is used to represent ontologies

- **Brief summary of benefits:**

Sambo allows users to integrate ontologies, this is a key functionality in bioinformatics as there are many ontologies, which partially overlap. For specific applications, users typically need a domain specific ontologies, which integrates parts of existing ontologies such GeneOntology, Mesh, UMLS, SnoMed, etc. Thus, Sambo is very important for an ontology-based information integration tasks.

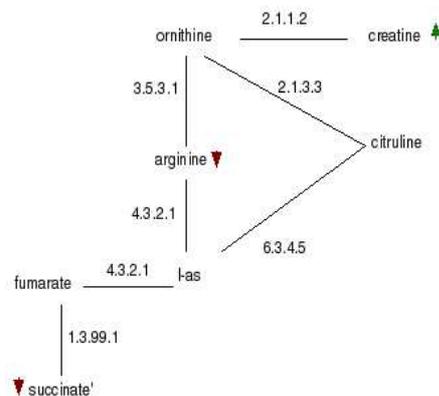


Figure 3: An example of a metabolic sub-network involving metabolites affected by hydrazine.

4 BioRevise

- **Name of the system:** BioRevise
- **Brief description of the system:** The BioRevise system is built on *REVISE* a non-monotonic reasoning system that revises extended logic programmes. It is based on logic programming with explicit negation and integrity constraints and provides two-valued revision assumptions to remove contradictions from the knowledge base. This Belief revision system is used to model inhibition of reactions in metabolic and biosynthesis pathways to explain the alternation on the concentration levels of metabolites after being exposed to external factors like toxins. The system will process the data related to the *Pathways* which are extracted from the *KEGG* pathways database.
- **Requirements:** Swiprolog installation and REVISE.
- **Example:** Identifying the effect of toxin in metabolic networks.

BioRevise can predict inhibitory effects of hydrazine on metabolic networks to assess the potential harmful side effects of a drug. Information on up/down changes in metabolite concentrations after hydrazine treatment is obtained from NMR spectra. This information is combined with KEGG metabolic diagrams, which contain information on the chemical reactions and associated enzymes (see Fig. 3).

The example is implemented as follows:

- Observables

The observations of the concentration level of the metabolites after the use of Hydrazine.

```
concentration('succinate',down).
concentration('creatine',up).
```

- Background knowledge.

The background knowledge describes the reactions taking place between the metabolites and the enzyme involved in the metabolic networks as explained in the KEGG database.

```
reaction(['succinate'], '1.3.99.1', ['fumarate']).
reaction(['fumarate','arginine'], '4.3.2.1', ['l-as']).
reaction(['l-as'], '6.3.4.5', ['citruline']).
reaction(['ornithine'], '2.1.3.3', ['citruline']).
reaction(['arginine'], '3.5.3.1', ['ornithine']).
reaction(['ornithine'], '2.1.1.2', ['creatine']).
```

– Abducible Predicate.

```
revisable(inhibited(_)).
```

The abducible indicates that the predicates truth-value can be changed in order to remove contradictions.

– Program rules.

The following rules describes the underlying mechanism of the effect of inhibition of a toxin by defining the observable concentration predicate.

```
concentration(Sub,down) <-
    reaction(Sub,Enz,Prod), % X <--Enz-- Y
    inhibited(Enz),
    concentration(Prod,up).

concentration(Sub,V) <-
    reaction(Sub,Enz,Prod), % X <--Enz-- Y
    not inhibited(Enz),
    concentration_all(Prod,V).

concentration(Sub,up) <-
    reaction(Prod,Enz,Sub), % Y <--Enz-- X
    inhibited(Enz),
    concentration(Prod,down).
```

– Integrity constraints.

```
<- concentration(M, up), concentration(M, down).
```

– Possible results for the above example

```
?- solution(A).
```

```
A = [[inhibited('1.3.99.1')], []] ;
```

A = [[inhibited('2.1.1.2')], []] ;

A = [[inhibited('2.1.3.3'), inhibited('3.5.3.1')], []] ;

A = [[inhibited('3.5.3.1'), inhibited('4.3.2.1')], []] ;

A = [[inhibited('3.5.3.1'), inhibited('6.3.4.5')], []] ;

No

- **Brief description how rules and other technologies are used:**

The whole system is modelled using Prolog rules: examples:-

reactionnode(Metabolite1, Enzyme, Metabolite2)

describe the topology of the network of the *metabolic pathways*. At the end we can get as a result, possible explanations for the change of the concentrations related to the observation, knowing which enzymes are affected and therefore which reactions are inhibited. This explanations are presented in predicates:

inhibited(Enzyme, Metabolite1, Metabolite2)

capturing the hypothesis that the reaction from *Metabolite1* to *Metabolite2* is inhibited by one toxin through an adverse effect on the enzyme, *Enzyme*, that normally catalyzes this reaction .

We need to take into consideration the predicate:

reactionnode(Metabolites1, Enzymes, Metabolites2)

where *Metabolites1*, *Metabolites2* and *Enzymes* could be a list of metabolites and enzymes, respectively.

- **Brief summary of benefits;**

In the development of new drugs it is useful to know all metabolic reactions affected by the drug and which enzymes will be inhibited with the introduction of toxins in the system. Since the number of *metabolic pathways* is huge, this work contributes explanations about how observed changes in the concentration of metabolites can be explained by inhibited enzymes in *metabolic pathways*.

5 EarthFeed and BioChemFeed

- **Name of the system:** *EarthFeed* and *BioChemFeed*

- **Brief description of the system:**

The EarthFeed engine is a RSS/ATOM News Feed visualization tool that displays news summaries in the context of their geographical localization. A high-resolution satellite image (210 mega-pixels) is used as background and provides a beautiful way to stay Earth-tuned. The system relies on the analysis of raw free text and its categorization against an ontological background knowledge. In EarthFeed, a straightforward ontology for continents, countries, cities, places of interest and institutions provides terms that are looked for and extracted from the title and summary of the news item. This simple geographical ontology admits an obvious and beautiful visualization, a high resolution satellite image of the earth itself. News are being fetched continuously from a list of URLs. The systems automatically moves all around the world, displaying news items as they come.

Building on the abstract idea of *Semantic Visualization*, we designed BioChemFeed: replace the earth image with a digital scan of the Biochemical Pathways (the so-called Bohringer map) and the ontology by the enzymes mentioned in the map. News items mentioning enzymes can simply be obtained through the new RSS facility offered by PubMed for a given search.

- **Requirements:**

- Java Runtime Environment 1.5;
- EarthFeed JAR file.

- **Brief description how rules and other technologies are used:**

RSS and ATOM standards are supported by EarthFeed and BioChemFeed. The system is reactive and hence could deploy reaction rules.

- **Brief summary of benefits:**

EarthFeed and BioChemFeed are experiments in semantic visualization, giving an example of how internet connectivity, ontological knowledge, and visualization can be combined together.

6 GoPubMed

- **Name of the system:** GoPubMed

- **Brief description of the system:**

GoPubMed is an alternative to classical literature search. GoPubMed submits keywords to PubMed, extracts GO-terms from the retrieved abstracts, and presents the induced ontology for browsing. The induced ontology is the minimal subset of GO, which comprises all GO terms found in the documents. The users actually queries PubMed.

- **Requirments:**

GoPubMed is a freely available Web application. A DSL internet connection and a modern browser will do.

- **Example:**

Example: Actin:

Which term is most obviously related to actin? Many researchers will promptly reply myosin. In GoPubMed such obvious relationships can be identified by exploring the most frequently occurring GO terms. In the case of actin GoPubMed suggests that some 80 papers mention cellular components or any sub-terms, nearly 80 papers cell or sub-terms, some 70 intracellular, 67 cytoplasm, 57 cytoskeleton, 50 actin cytoskeleton and 9 myosin. Thus, in only 5 clicks the user can relate actin and myosin and even underpin this relationship through the statements of associated abstracts, such as PMID 15679101: Syntrophin was also able to inhibit actin-activated myosin ATPase activity.

Example: Author profiles

GoPubMed is generally useful to gain an overview over a set of articles and to define a profile for these articles. This feature can be used to quickly get an insight into the topics a researcher is working on. Specifying e.g. the name and affiliation of a researcher as query to GoPubMed one will be able explore the researcher's interest and focus of research. In particular, the induced GeneOntology can serve as a profile representing that researcher. As an example, consider Kai Simons in Dresden. The PubMed query simons dresden returns some 20 articles. The induced ontology for these papers indicates that he is working on cell organisation and biogenesis (within the process ontology) and in particular on lipid raft formation, a term that is found in 13 papers.

- **Brief description how rules and other technologies are used:**

GoPubMed handles ontologies in OWL and documents in PubMed XML. GoPubMed implements simple reasoning computing induced ontologies from given documents.

- **Brief summary of the benefits:**

First, it gives an overview over literature abstracts by categorizing abstracts according to the Gene Ontology and thus allowing users to quickly navigate through the abstracts by category. Second, it automatically shows general ontology terms related to the original query, which often do not even appear directly in the abstract. Third, it enables users to verify its classification because GeneOntology terms are highlighted in the abstracts and as each term is labelled with an accuracy percentage. Fourth, exploring PubMed abstracts

