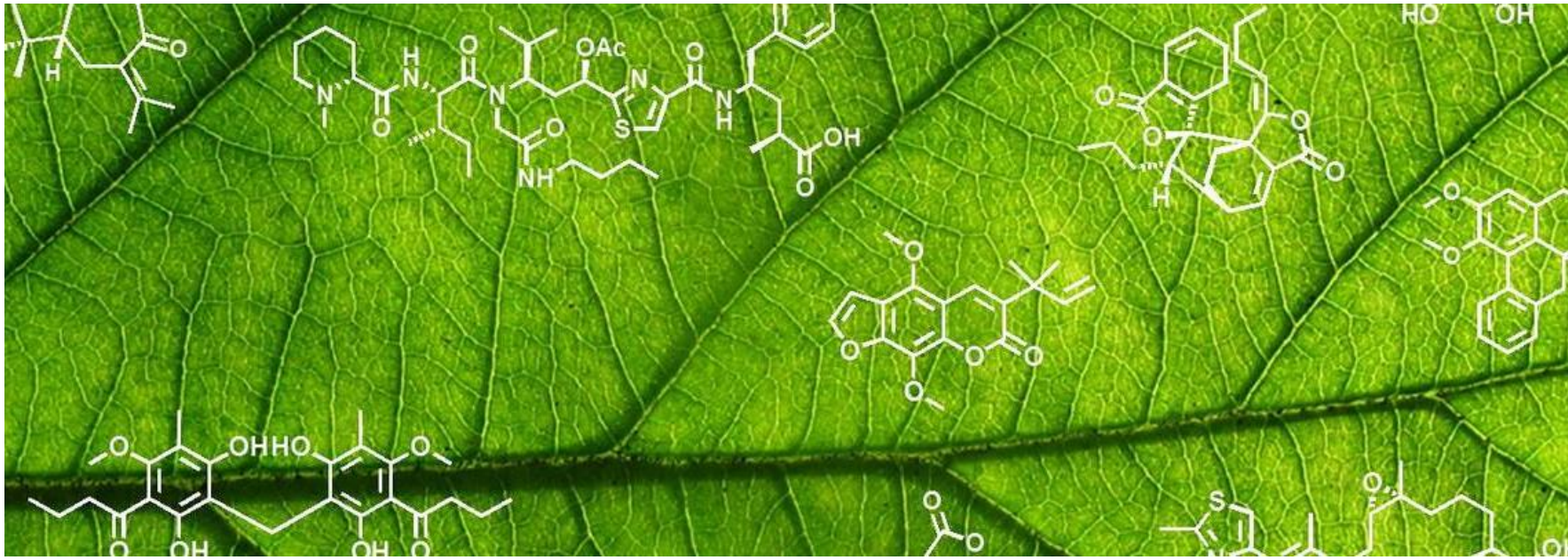
IPB Leibniz-Institut für Pflanzenbiochemie



Concept of the Leibniz Bioactives Cloud

Frank Broda & Frank Lange

Leibniz Bioactive Compounds and Biotechnology

Leibniz-Gemeinschaft

WissenschaftsCampus Pflanzenbasierte Bioökonomie Halle

LRA Bioactive Compounds and Biotechnology

- started in 2013; now 19 partner institutes from sections C, D & E

- tasks:
  - combine expertise to solve existing and upcoming challenges in bioactive compound development
  - initiate strategic partnerships
  - allow and ensure funding of relevant research programs

- focus groups:
  - collections and infrastructure
  - pharmaceutical agents
  - non-medical bioactive agents
  - biotechnology and sustainable production
  - translational research and toxicology

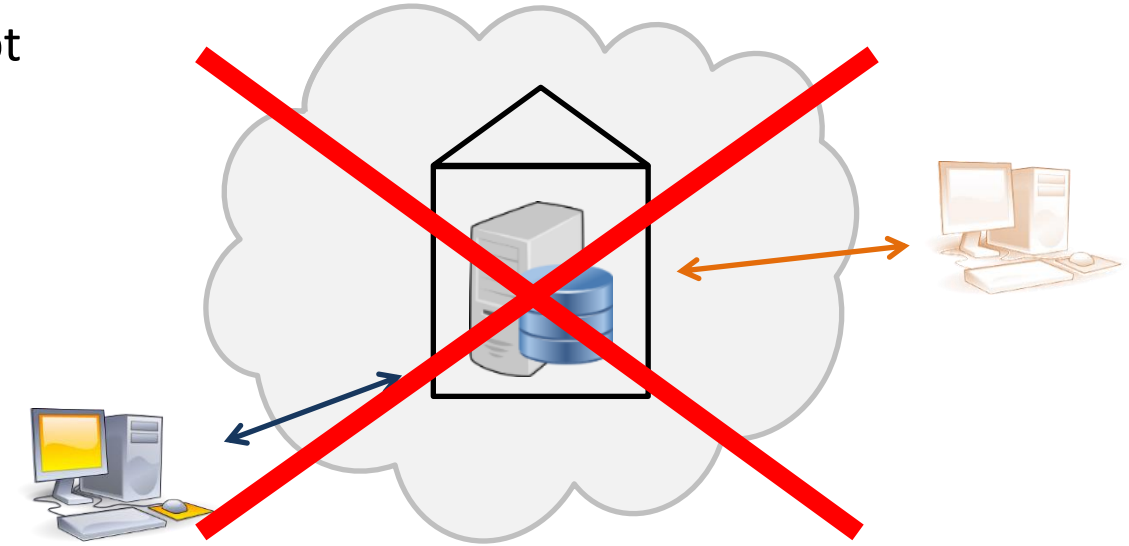
- public appearance: organizes the Leibniz Conference on Bioactive Compounds (Wirkstofftage) and awards the Leibniz Drug of the Year

Purpose of the Leibniz Bioactives Cloud

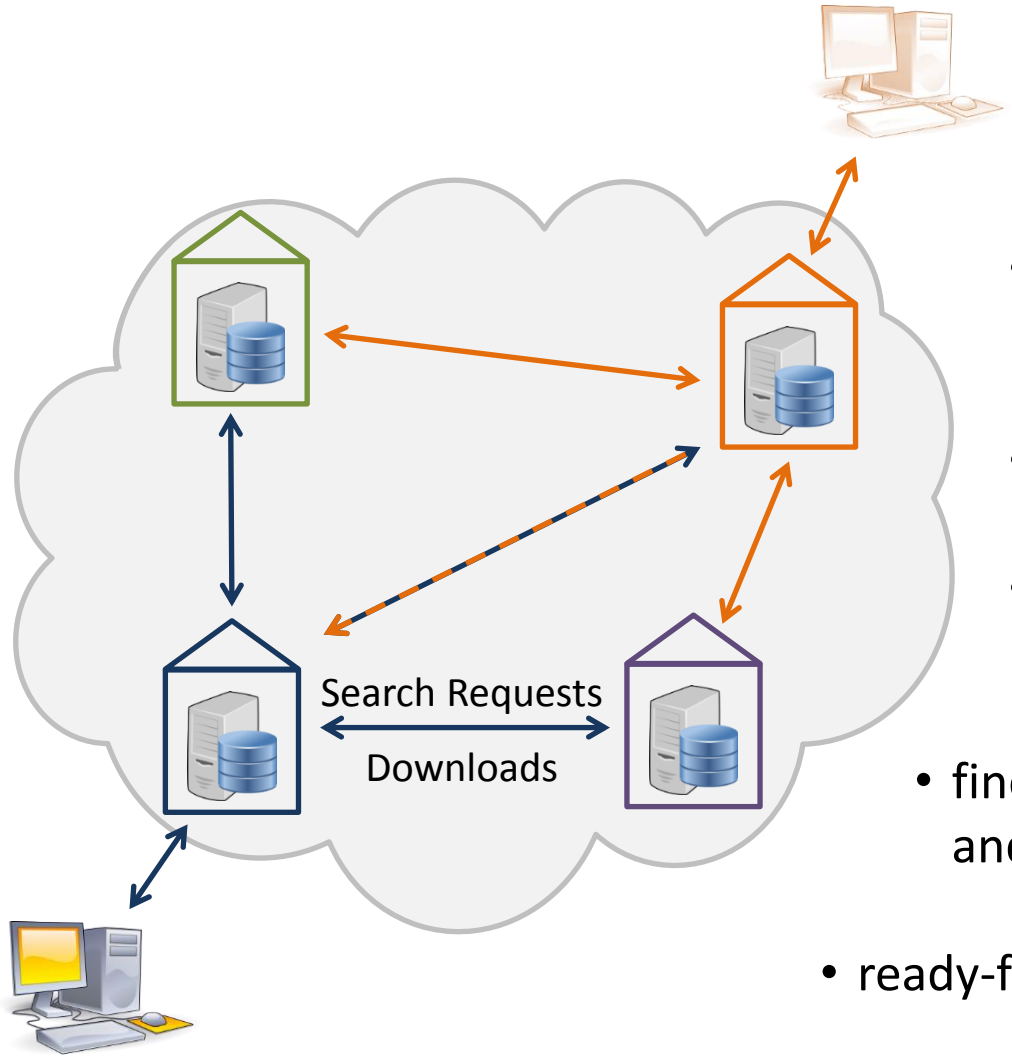
- provide a “smart data warehouse” as exchange portal for the members of the LRA

- What to exchange?
  - text documents
  - research data (e.g. results of assay experiments, compound libraries)

- no “centralized cloud” concept



General Structure

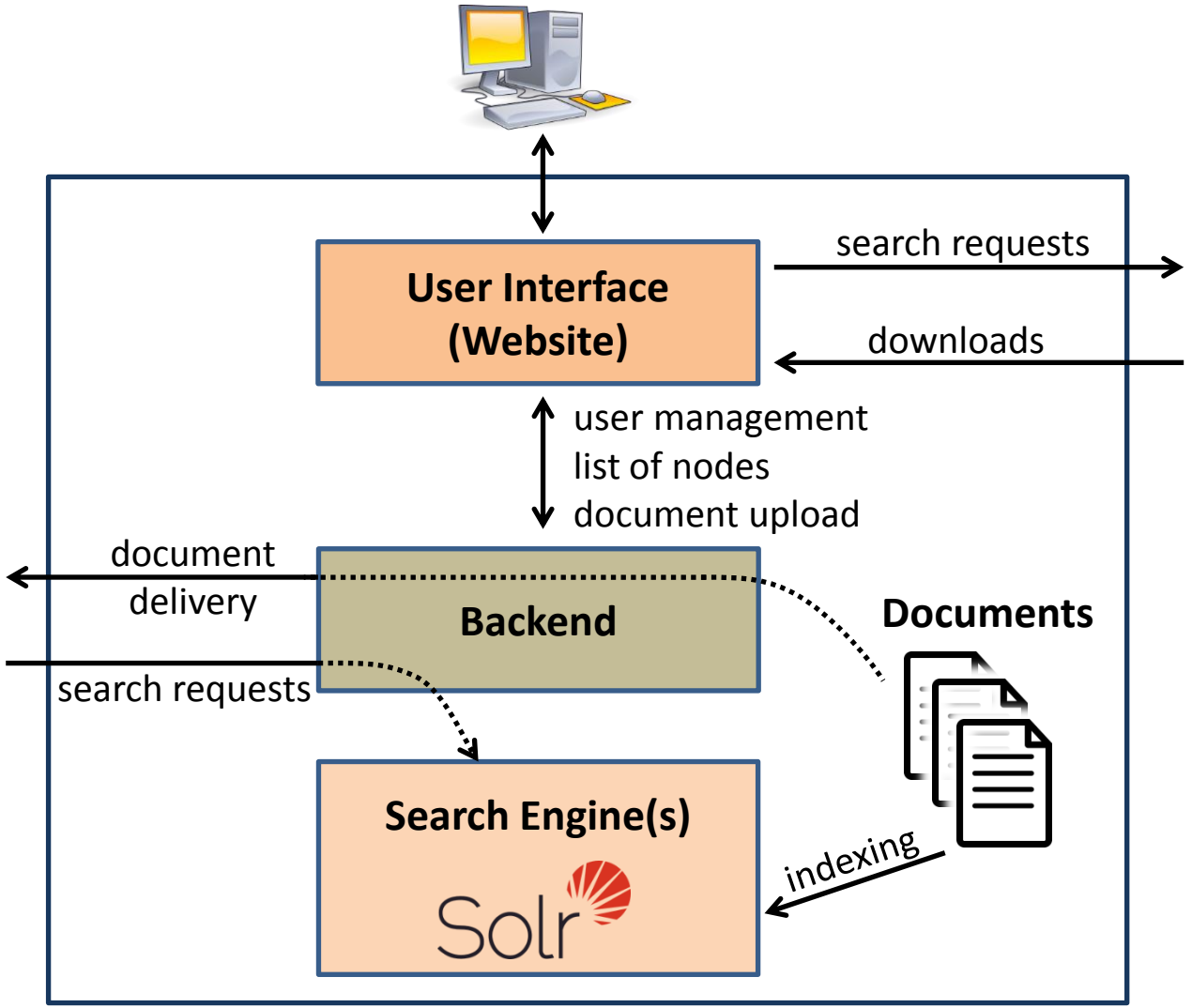


- distributed infrastructure with nodes in participating institutes
- interaction via user interface
- stored data remains at the respective institute
- fine-grained access control for search and document download
- ready-for-use software (virtual machine)

General Structure

Design of a node

=



Challenges

Distributed infrastructure:

- search function throughout the nodes (= distributed/federated search)
  - no out-of-the-box feature of Solr
  - issues with scoring/ranking of search results

⇨ learn from BioSolr project

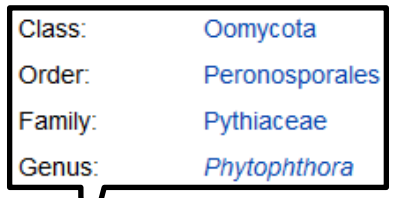
Becoming “smart” - a semantic search engine

- focus on text documents first
  - research & project reports
  - article(-preprints)
  - theses

⇨ **text mining with ontologies**

Text Mining with Ontologies

- mining strategy includes indexing with ontologies
  - organisms
  - extracts
  - assays
  - compounds
  - reactions
  - etc.



BIOACTIVE FUNGAL METABOLITES

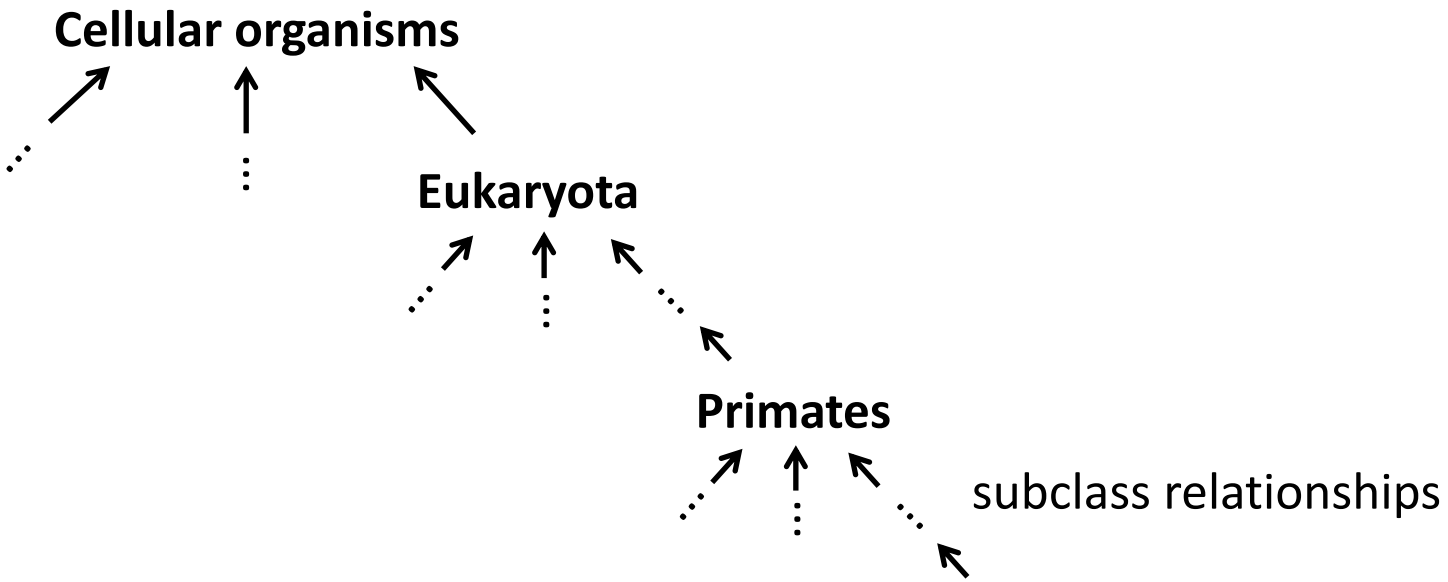
Infections with *Phytophthora infestans*, the causal agent of potato and tomato late blight disease, are difficult to control and can lead to considerable agricultural losses. Thus, the development of new effective agents against the pathogen is of great interest. In previous work, (*E*)-4-oxohexadec-2-enoic acid (**compound 1**) was isolated from *Hygrophorus eburneus*, which exhibited fungicidal activity against *Cladosporium cucumerinum*. In our on-

Class: 4-Oxo-2-alkenoic Fatty Acids

⇨ **Search: Oomycota AND fatty acid**

Scientific Report 2009-2010, IPB

An ontology for organisms: NCBI Taxonomy



unique identifier: NCBITaxon_9606
Synonyms: “human”, “man”, “humans”

Text Mining with Ontologies

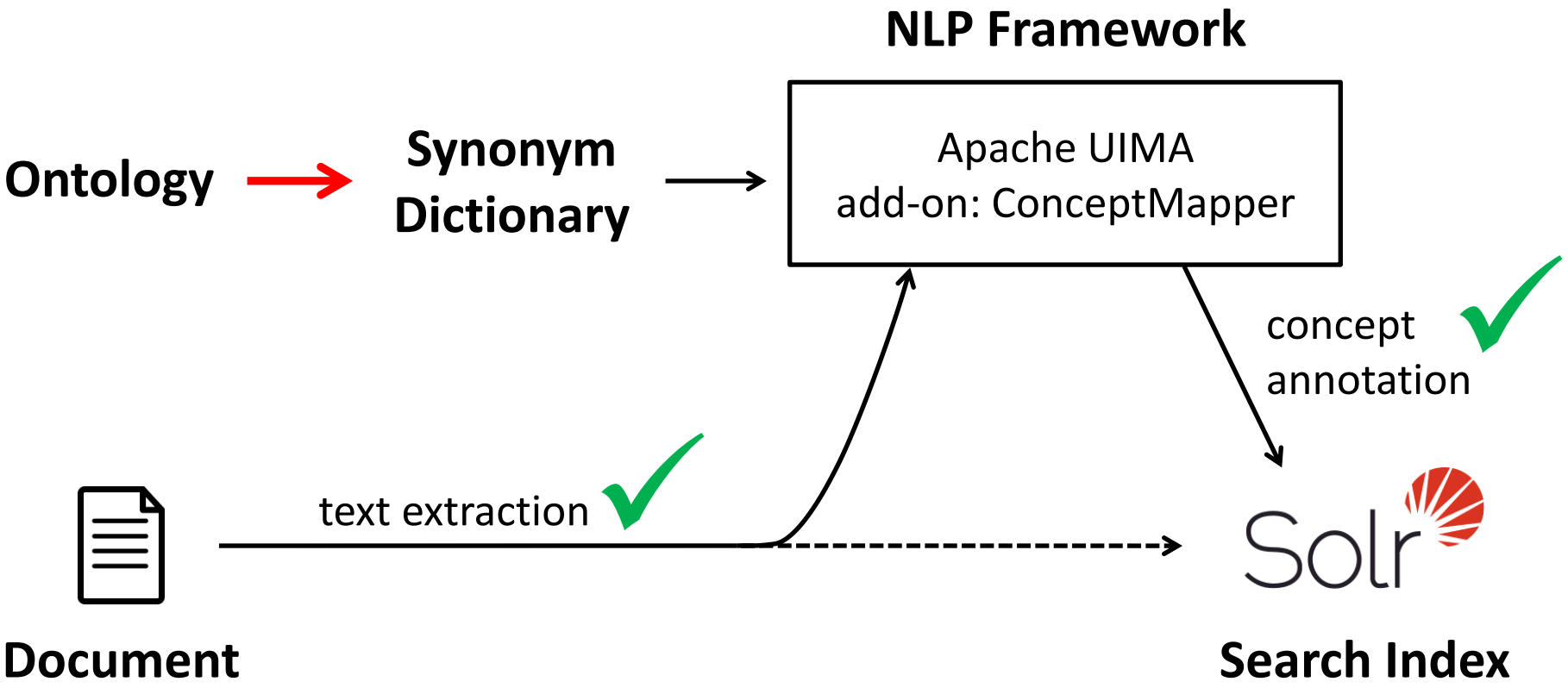
from ncbitaxon.owl:

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/NCBITaxon_9606">
 <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/NCBITaxon_9605"/>
 <ncbitaxon:has_rank rdf:resource="http://purl.obolibrary.org/obo/NCBITaxon_species"/>
 <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">GC_ID:1</oboInOwl:hasDbXref>
 <oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/XMLSchema#string">human
 </oboInOwl:hasExactSynonym>
 <oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/XMLSchema#string">man
 </oboInOwl:hasExactSynonym>
 <oboInOwl:hasOBONamespace rdf:datatype="http://www.w3.org/2001/XMLSchema#string">ncbi_taxonomy
 </oboInOwl:hasOBONamespace>
 <oboInOwl:hasRelatedSynonym rdf:datatype="http://www.w3.org/2001/XMLSchema#string">humans
 </oboInOwl:hasRelatedSynonym>
 <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Homo sapiens</rdfs:label>
</owl:Class>
```

Strategy:

- use synonyms for text mining
- annotate identifier in search index

Text Mining with Ontologies



Searching with Ontologies

Search Query: “Primate” —map with ontology→ NCBITaxon_9443

Query expansion: also consider all subclasses of NCBITaxon_9443

NCBITaxon_9443 (Primates)
NCBITaxon_376913 (Haplorrhini)
NCBITaxon_376911 (Strepsirrhini)
…
NCBITaxon_9593 (*Gorilla gorilla*)
NCBITaxon_9606 (*Homo sapiens*)

Solr

Document list

Searching with Ontologies

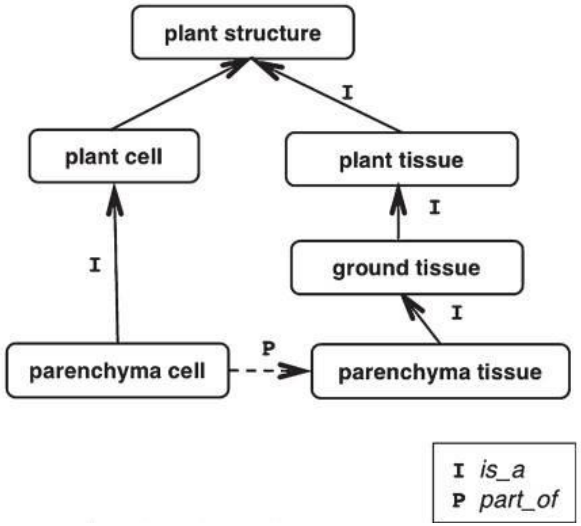
Problems:

- search query will explode (NCBITaxon_9443 has 868 subclasses)

⇨ performance issues?

- scoring/ranking of results will be messed up

- ontologies can be more complex
  - not tree-like; cyclic graphs (no hierarchy)
  - more than “subClassOf” relations

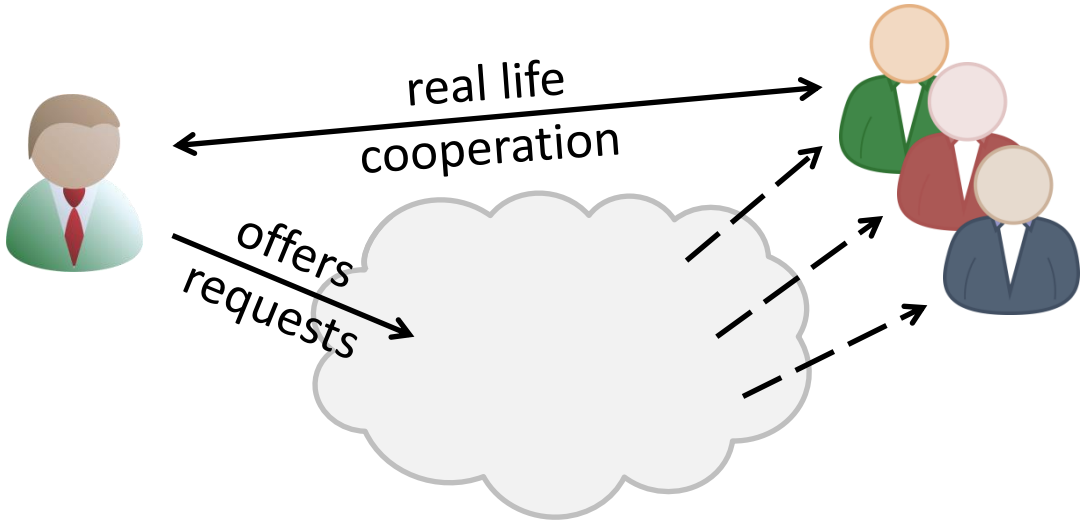


Walls et al. *Am. J. Bot.* **2012**, 99, 1263-1275

The Bioactives Cloud as active communication platform

- help initializing new cooperations among the members of the LRA

  use case: bring compound providers and compound testers together



⇨ black board / ticket system
⇨ semantic annotations

Backup slides

TF-IDF scoring model

Term Frequency (*TF*)

$$TF = \sqrt{f}$$

f: number of times the query term occurs in the document

Inverse Document Frequency (IDF)

$$IDF = 1 + \log\left(\frac{numDocs}{docFreq + 1}\right)$$

numDocs: number of documents in corpus

docFreq: number of documents with the query term

⇨ **Score = *TF* * *IDF***

Headline

This is the template slide
Please don’t delete it!

Citation