Combining Fact and Document Retrieval with Spreading Activation for Semantic Desktop Search

Kinga Schumacher, Michael Sintek and Leo Sauermann

{firstname.surname}@dfki.de

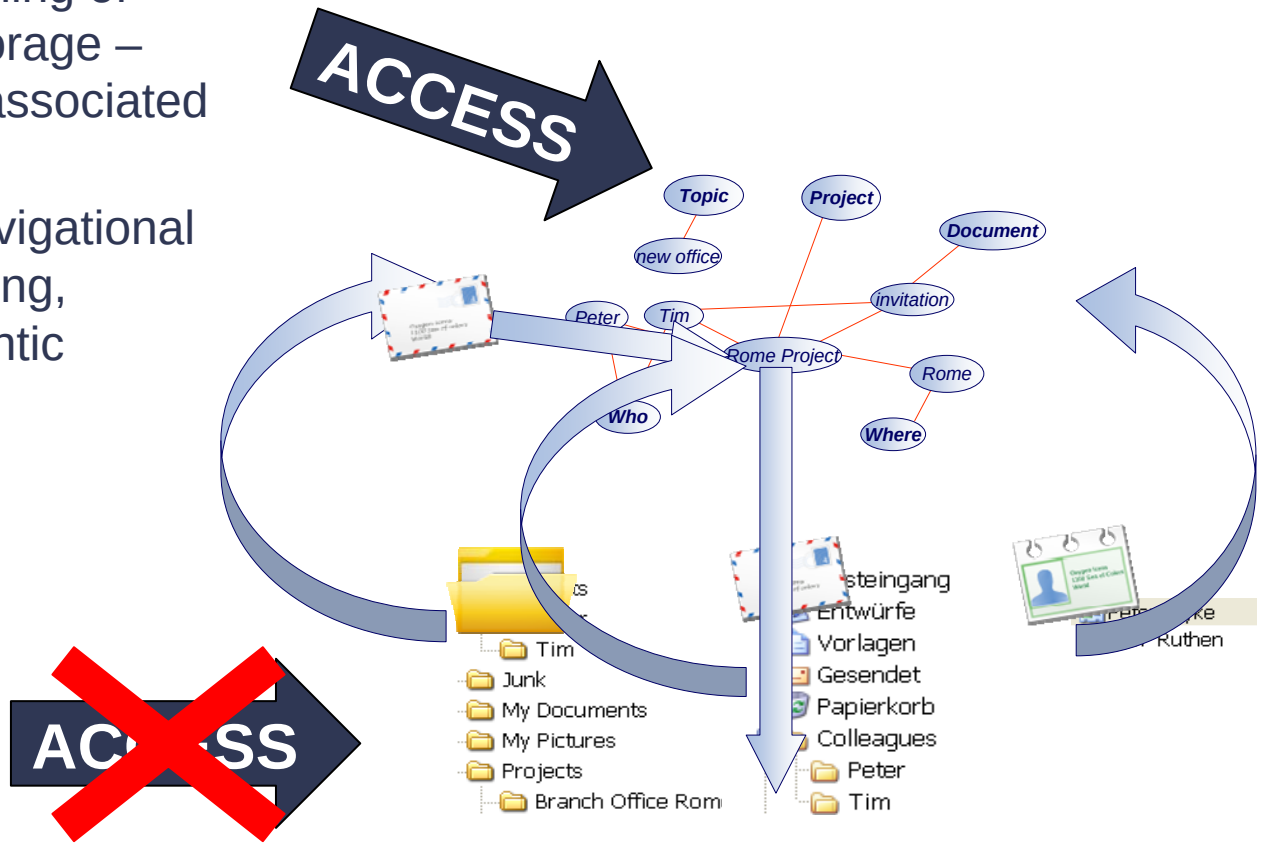
German Research Center for Artificial Intelligence (DFKI GmbH)

Outline

- Semantic Desktop
- Semantic Search research areas
- Our approach
- Evaluation
- Future work

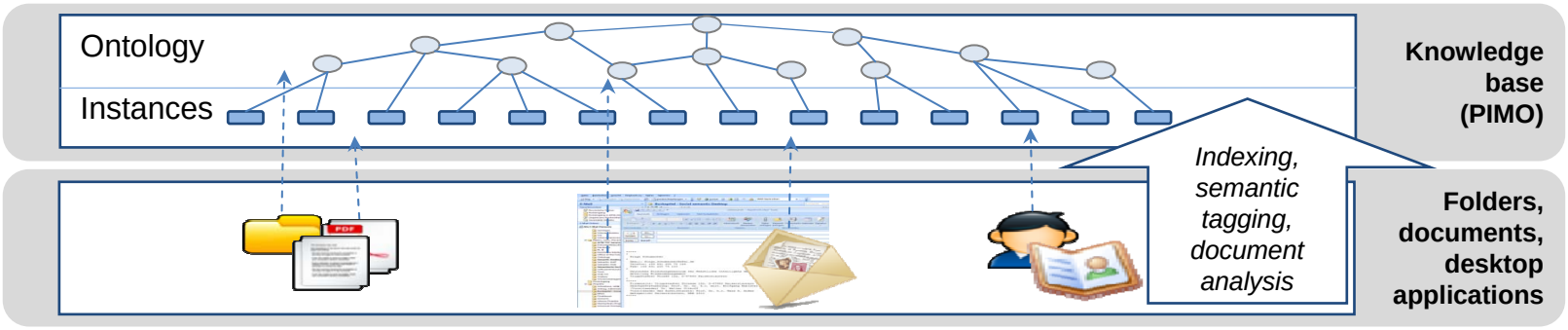
- Means for ***Personal Information Management***
- RDF, RDFS, identification of resources by URIs
- Instead of a document- and application-oriented information management, the Semantic Desktop enables the user to
  - create own categorization system of projects, persons, topics, events, locations, organizations etc.
  - integrate all resources (e.g. text-documents, contacts, messages, multimedia) across application borders
  - collect facts about them
  - annotate, classify and relate them building the **P**ersonal **I**nformation **M**odel (PIMO)

- Supports the user with
  - Keeping: handling of information storage – concepts are associated with folders
  - Finding: by navigational search, browsing, filtering, semantic search



Information – the knowledge base

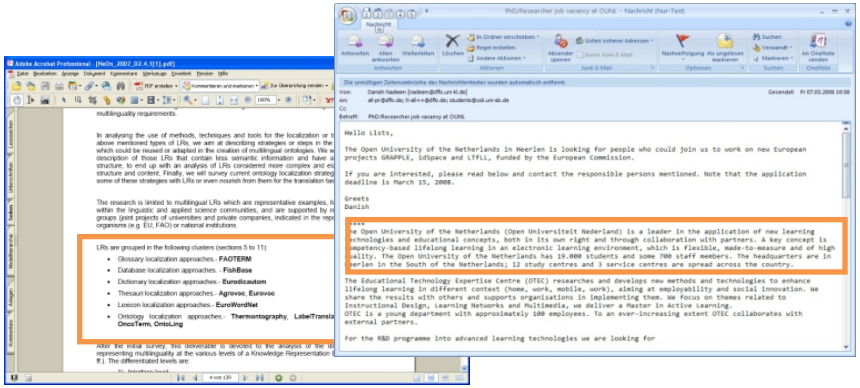
- Structured and unstructured: facts and documents
  - native structures (file system, email folders) are mapped to ontological concept
  - files and other information objects like contacts, calendar entries are mapped to instances
  - their textual content is indexed
- in ontologies, instance base and document-index



Human Access

- Search for Information: documents and facts

“seminar topics”



“phone number of the KM-Group secretary”

+49 631 205 75 101

- Enable Free-text queries
  - to keep knowledge overhead away from the user
  - NLP problems, e.g. syntactic, structural ambiguity

Semantic *Document* Retrieval

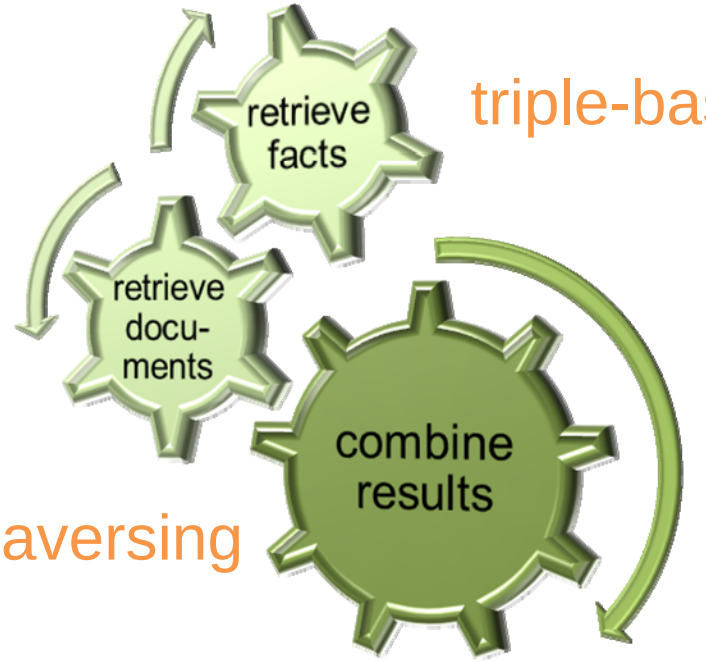
Document retrieval techniques enhanced through

- usage of linguistic information
- usage of category systems
- graph traversing

Fact Retrieval

Fact retrieval through

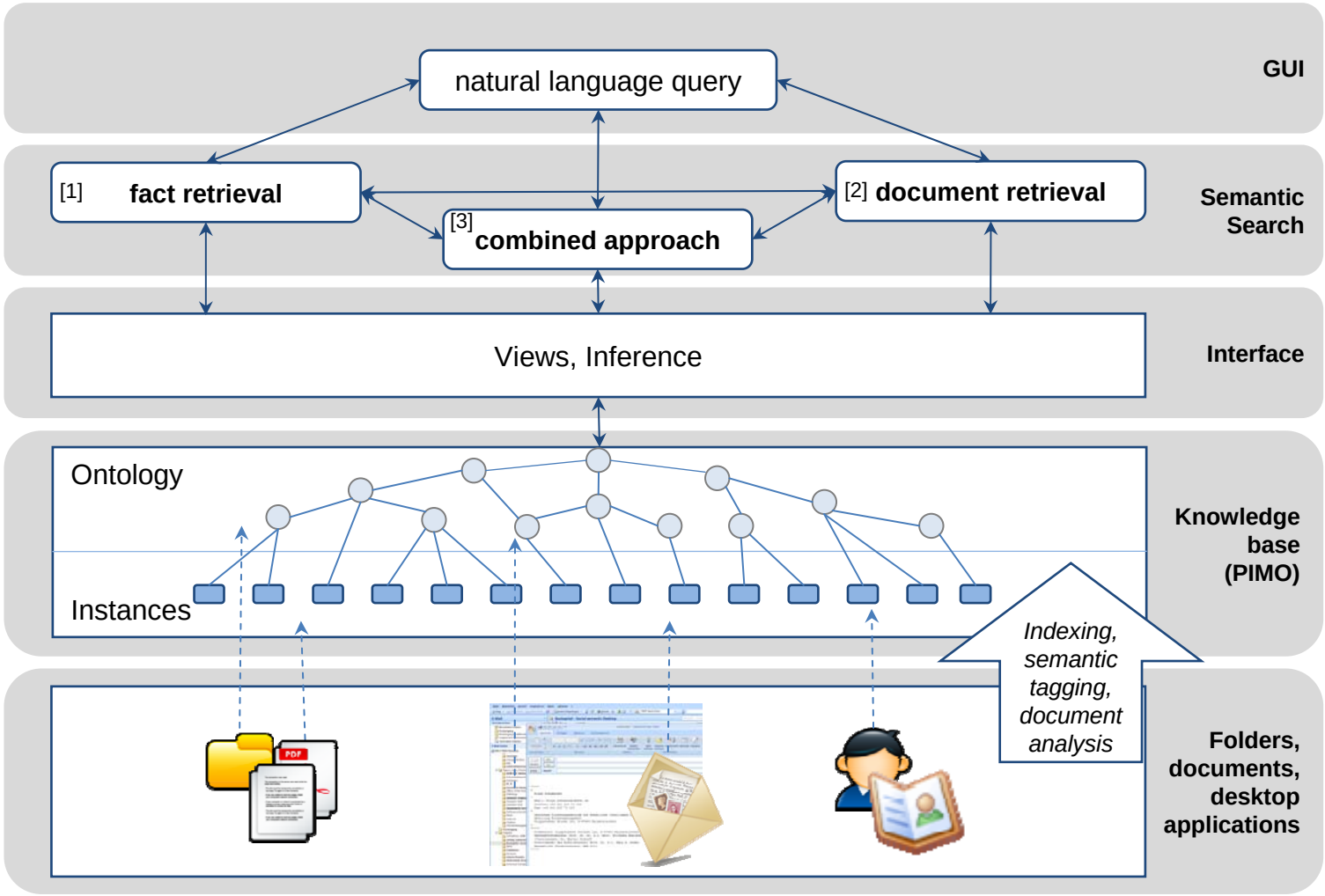
- reasoning
- triple(statement)-based algorithms
- graph algorithms



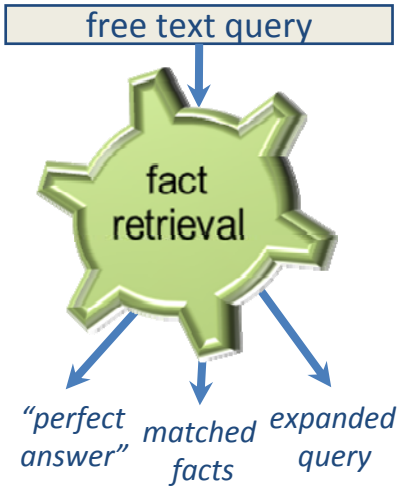
triple-based

graph traversing

Architecture



Fact Retrieval – Triple-based approach



1. Syntactic Matching: query $\{t_1, t_2, ..., t_n\}$
   - linguistic information in the knowledge base
   - n-gram method
   - phrase matching

   Result: set of potential Properties p_i, Instances i_j, Classes c_k
2. Semantic Matching on the instance base (based on [1])

   1st level: create and apply query templates with the matches adjacent terms $(i_j, p_i, ?)$, $(i_j, ?, c_k)$, $(?, p_i, ?)$, …

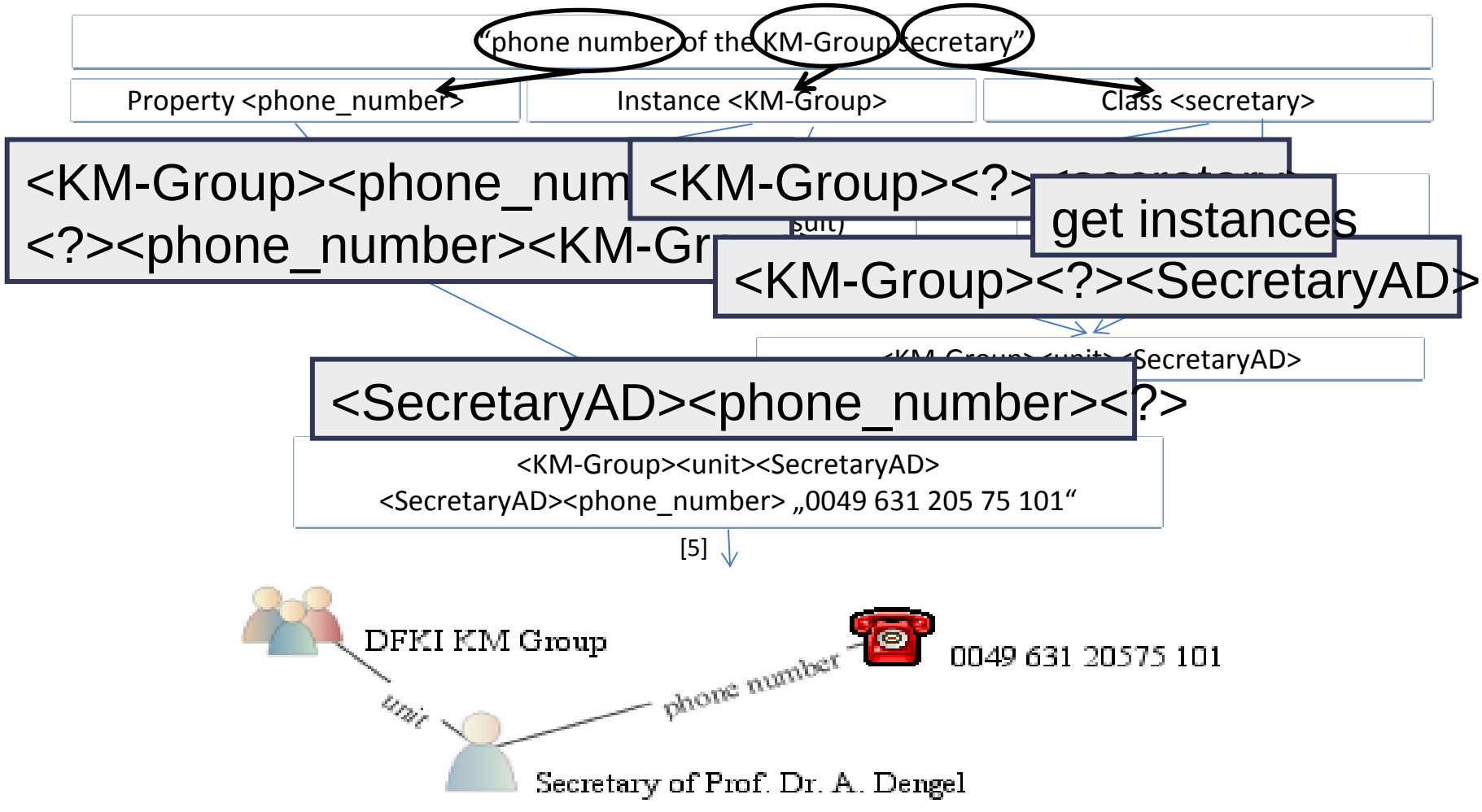
   2nd level:
   - iterate over found triples and the syntactic matches of until now semantically unmatched terms and create and apply query templates
   - stop when: all query terms are included or no further triples can be found

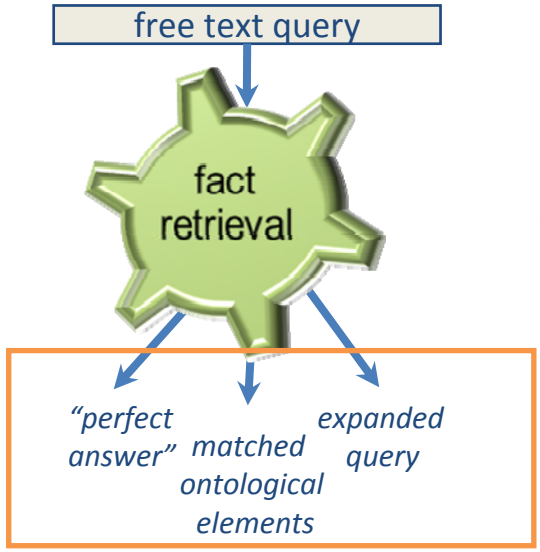
   3rd level:

   Combine found triples and identify result graphs (coherent subgraphs)

[1] D.E. Goldschmidt, M. Krishnamoorthy: Architecting a Search Engine for the Semantic Web. C&O-2005, Pittsburgh

Fact Retrieval Example





Results

Ranking

1. Syntactic Matching: n-gram weights $w(p_i), w(i_j), w(c_k)$
2. Semantic Matching:

1st level:

$$rank(< i >< p_i >< i_j >) = w(p_i) + w(i_j)$$

2nd level:

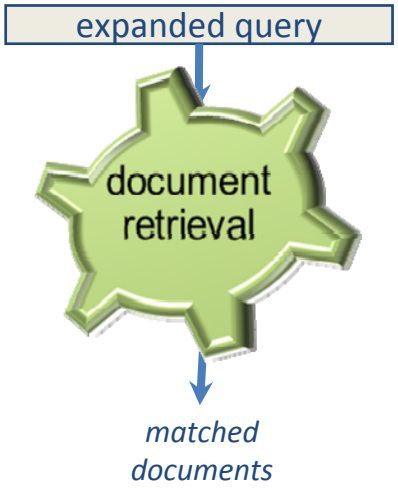
$$rank(triple\_set) = \sum(p_i) + \sum w(i_j) + \sum w(c_k),$$

where p_i, i_j, c_k are included in the triples

3rd level:

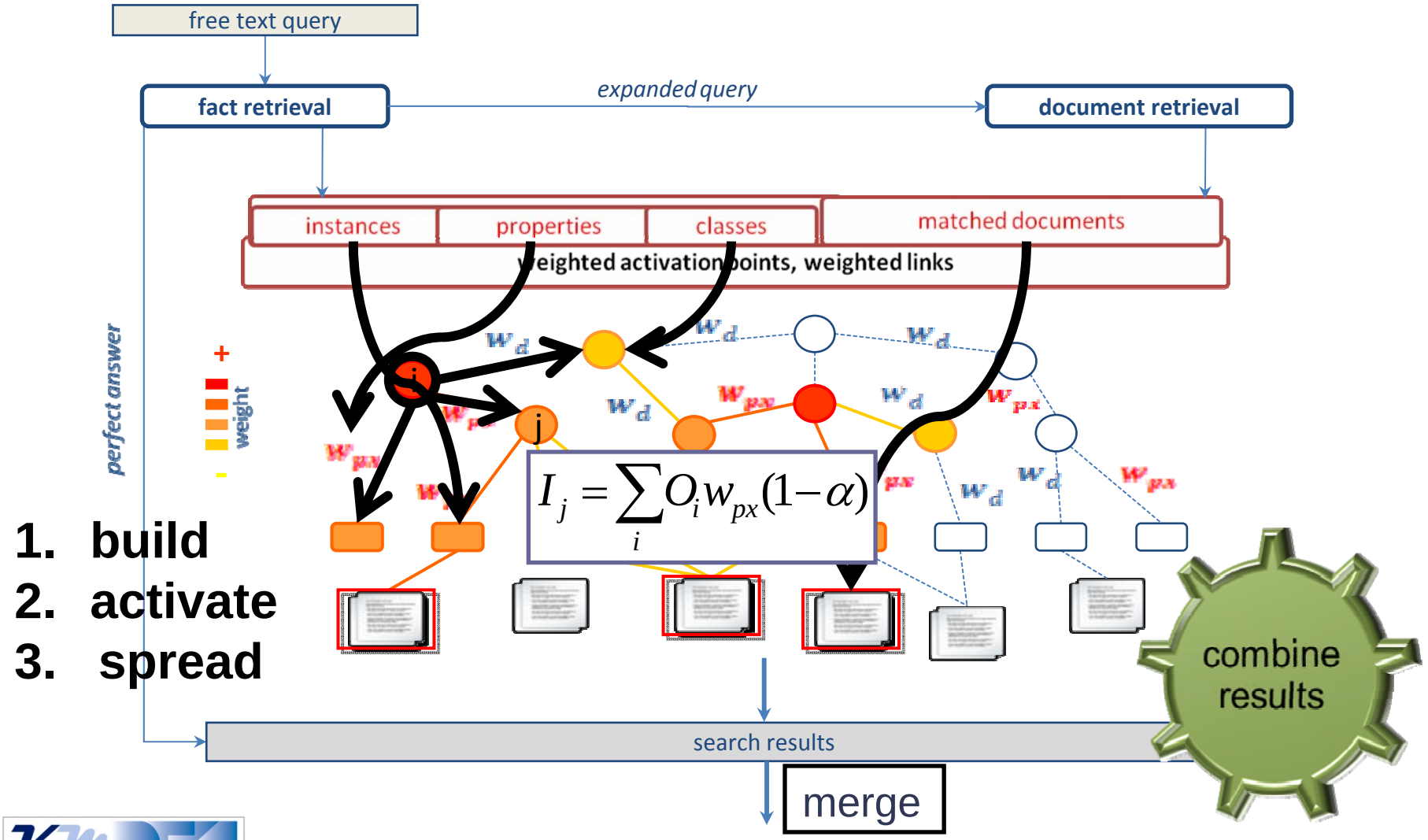
$$rank(subgraph) =$$
$$rank(triple\_set)/number\_of\_query\_terms$$

Semantic Document Retrieval – Graph traversing



- Expanded query: expanded with the linguistic information about the matched ontological elements
- Semantic Document Retrieval
  1. Keyword search on the document index (Lucene)
  2. Apply Spreading Activation:
     - Activation points: found documents
     - Activation weights: document weights
     - Formula: $I_j = \sum_i O_i w_{ij} (1-\alpha)$

Combined approach



presentation-1001 (KeynoteTalk)

hasRelatedDocument Emerging Sciences of the Internet: Some New Opportunities (Paper) (InProceedings)

author Ron Brachman (Person)

homepage http://isweb.uni-koblenz.de

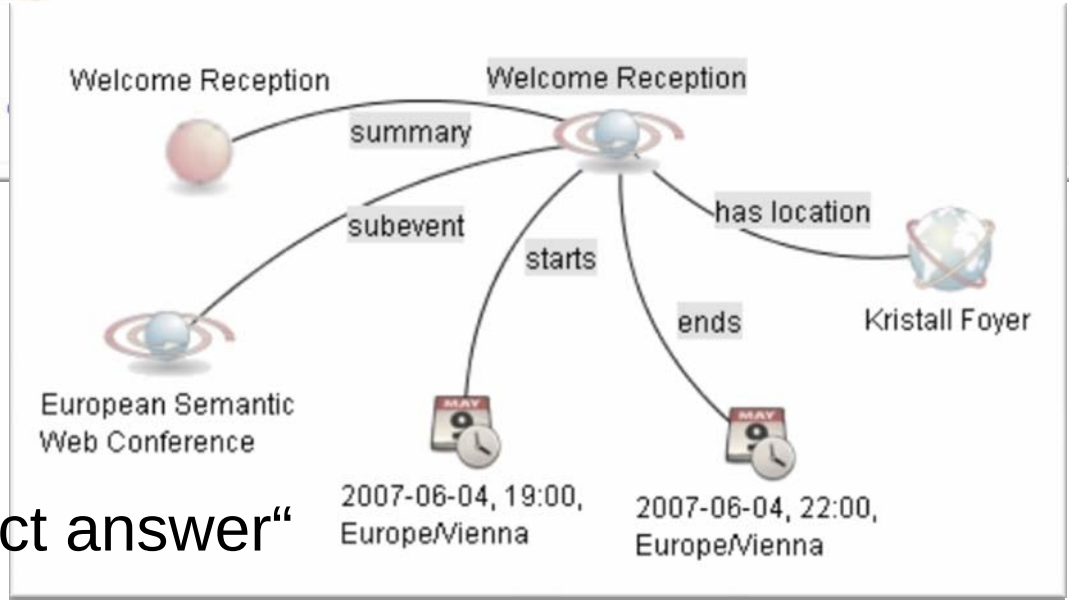
hasLocation Innsbruck (Place)

isSubEventOf Invited Talk: Ron Brachman (PaperSession)

hasLocation Innsbruck (Place)

isSubEventOf European Semantic Web

merged result



„perfect answer“

Data and method

- Standardized and annotated test data set for semantic desktop missing
- Evaluated with the ESWC 2007 knowledge base
- Knowledge base extended with some synonyms
- Evaluated against the Google Site search on www.eswc2007.org
- Set of 11 queries – typical queries of knowledge workers
- Average Precision (for details see Proceedings, pp 569-583)

	Semantic Desktop Search	Google Site Search
Average Precision	0.9436	0.4615

Strengths and Weaknesses

- ✓ precise results for complex queries
- ✓ recognition of phrases, synonyms
- ✓ resolving structural ambiguity
- ✓ enhanced ranking
- ✓ usef
  infor

- ✘ Lower precision by unsuitable long queries (if no properties matched: spreading activation propagates to all connected nodes with the same intensity)

need of more specific and personalized setup of the semantic network's link weights

-learn from feedback

-exploit context

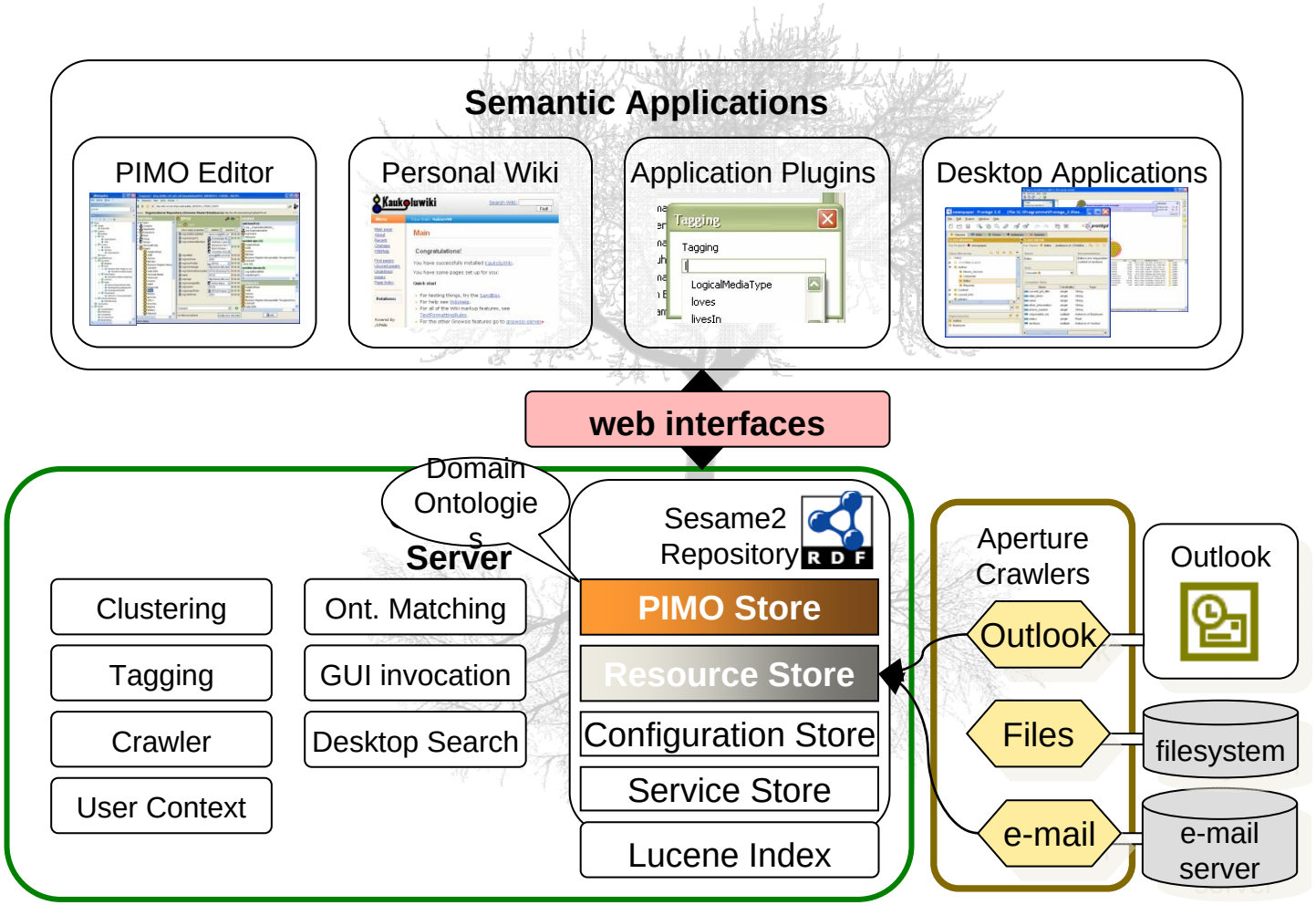
Future Work

- Gold Standard for Semantic (Desktop) Search Evaluations (in progress)
- Application of named graphs and views (based on the Nepomuk Representation Language NRL)
- Advanced GUI with dynamic filters and browsing support

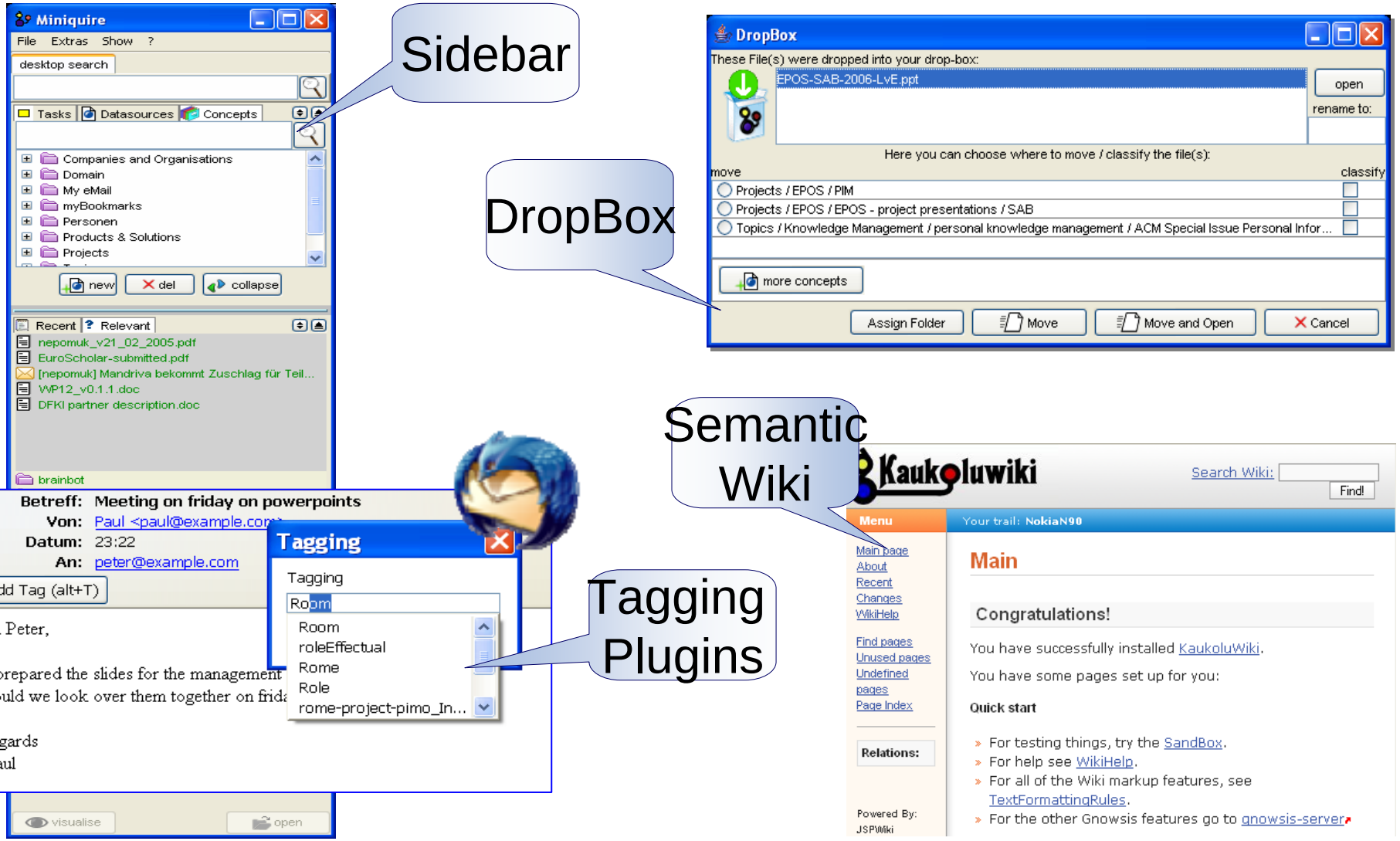
Thank you for your attention!

Thanks for the members of the DFKI KM-Group

Semantic Desktop Architecture



Semantic Desktop Tools



Extract of a PIMO

http://dengel.dfki.de/outlook/contact/0019E1773D6A56

has-address

http://www.mbwjk.rlp.de/ministerium/staatssekretaerin-dzwonnek.html

http://www.dfki.de/~wahlster/

http://www3.dfki.uni-kl.de/agd/dengel/content/index_ger.html

http://ag47.informatik.uni-bremen.de/eng/team.php

Calender (…)

takes-place-on

has-speaker

has-speaker

has-speaker

http://www.karl-rauber.de/

is-hosted-by

has-location

http://www.landesvertretung.rlp.de/haus_bruessel.html

file:///D:/EigeneDateien/Dokumente/Einladungen/Brüssel.pdf

is-hosted-by

is-hosted-by

http://www.landesvertretung.bremen.de/landesv/bevoll/index.html

- decompose a string in a subsequences of n characters

  ‚basic‘: ‚ba‘, ‚as‘, ‚si‘, ‚ic‘

  ‚base‘: ‚ba‘, ‚as‘, ‚se‘

- map the decomposition to a vector containing the number of occurrences of the n-grams

	ba	as	si	ic	se
basic	1	1	1	1	0
base	1	1	0	0	1

- compute the distance of the vectors

  e.g. Dice-Measure d(‘basic’,’base’)= 0.571