User Centered and Ontology Based Information Retrieval System for Life Sciences

Sylvie Ranwez
Mohameth-François Sy
Jacky Montmain
Michel Crampes

Vincent Ranwez

LGI2P Research Center / Ecole des Mines d'Alès, France ISEM – CNRS / Montpellier II University, France

Context and objectives

Ontology based information retrieval

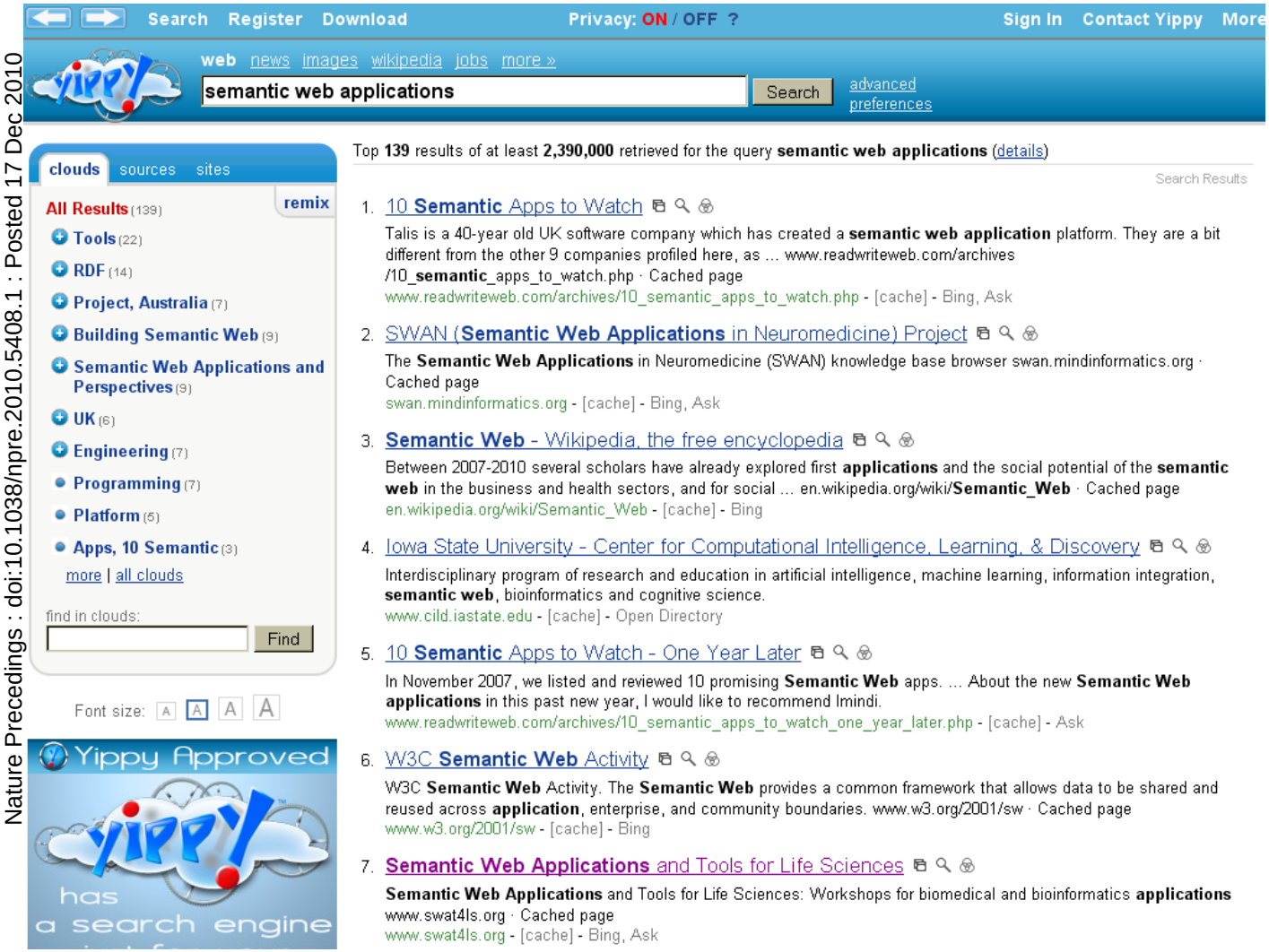
Relevance calculus between a document index and a query

- *Similarity between two concepts*
- *Relevance of a document with respect to a concept*
- *Relevance of a document with respect to a query*

Results visualization

Conclusion et perspectives

Context: *usual information retrieval engine*



Boolean search

+Results are easy to understand

-Exact terms matching

-Number of results

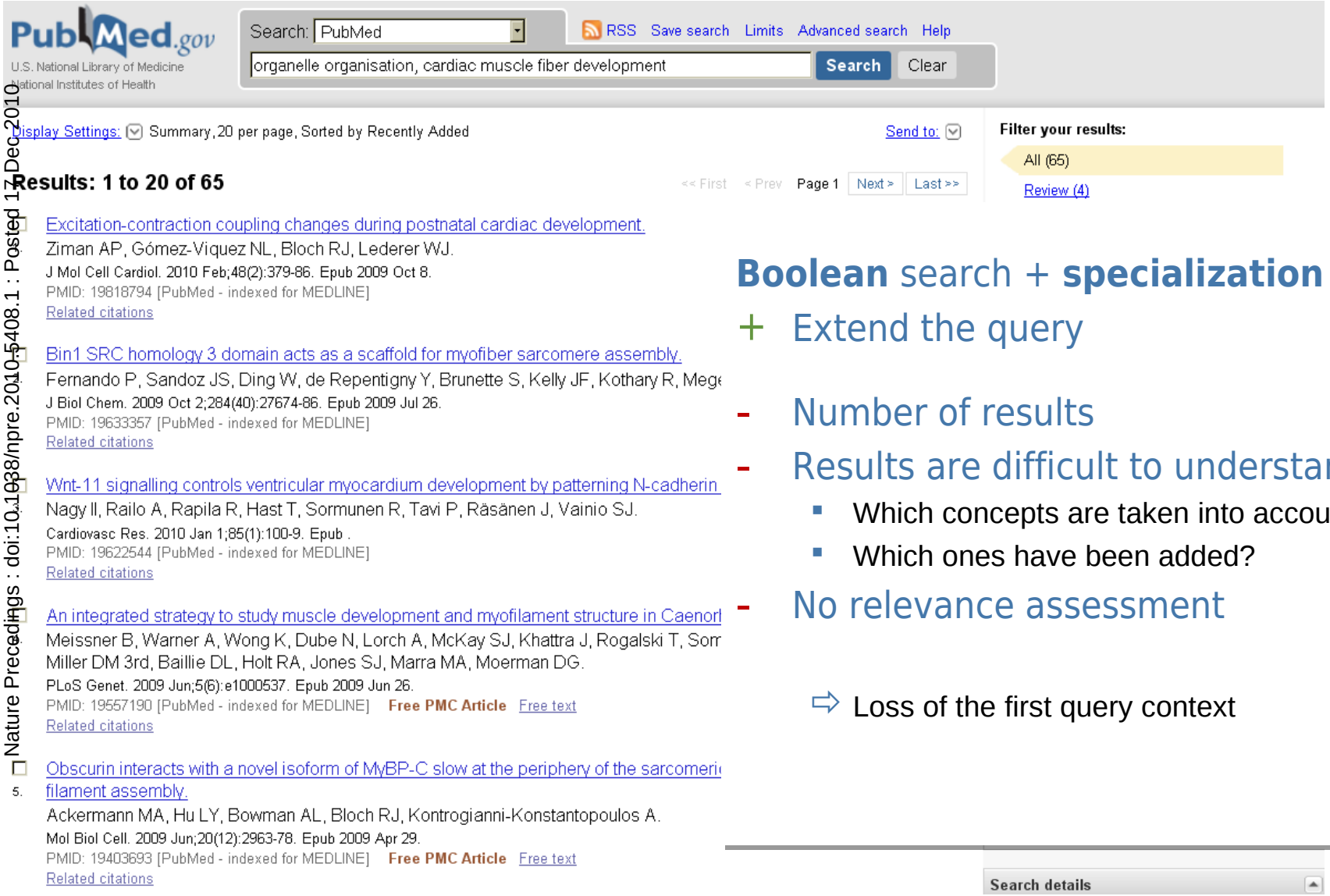
-Rough measurement: "match" or "does not match"

-Limited interaction

-Aggregating operators are not used (AND, OR...)

⇨ Hard to grasp even with clustering

Context: *information retrieval based on a concepts hierarchy*



Boolean search + **specialization**

\+ Extend the query

\- Number of results

\- Results are difficult to understand

- Which concepts are taken into account?
- Which ones have been added?

\- No relevance assessment

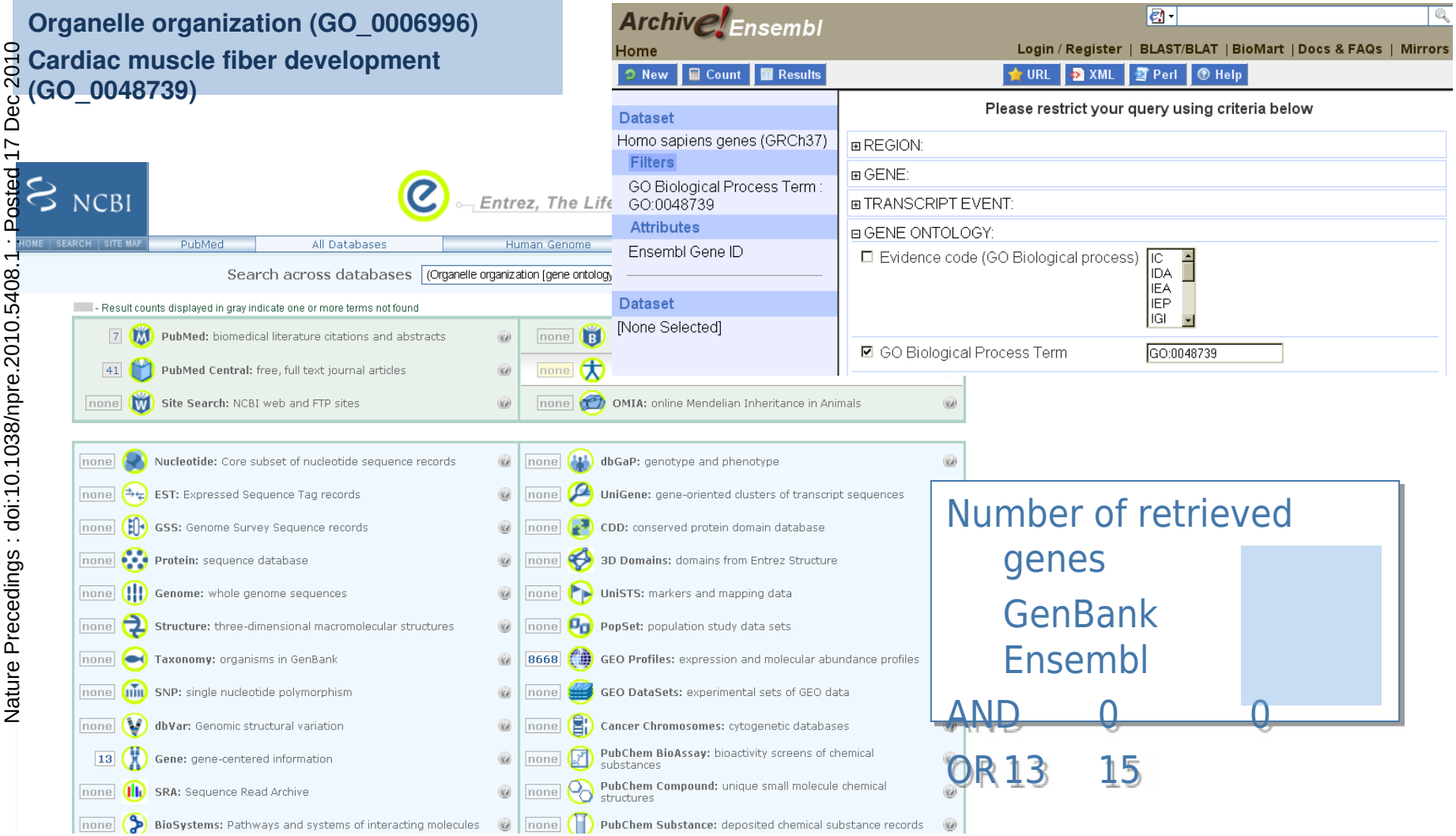
⇨ Loss of the first query context

Context: *information retrieval using ontologies*

?

Organelle organization (GO_0006996)

Cardiac muscle fiber development (GO_0048739)



Number of retrieved genes	GenBank	Ensembl
AND	0	0
OR	13	15

Take better benefits of ontologies during the information retrieval process (indexing/query matching)

- Expand the query if necessary
- Measure document/query adequacy by identifying added concepts

Favor the overall results' grasp by the user

- Explain why a document has been selected
- Give an overall vision of results
- If a selected document is not relevant, identify why in order to reformulate the query conveniently

Taking user preferences into account

⇨ Favor interactions and iterative querying process

Context and objectives

Ontology based information retrieval

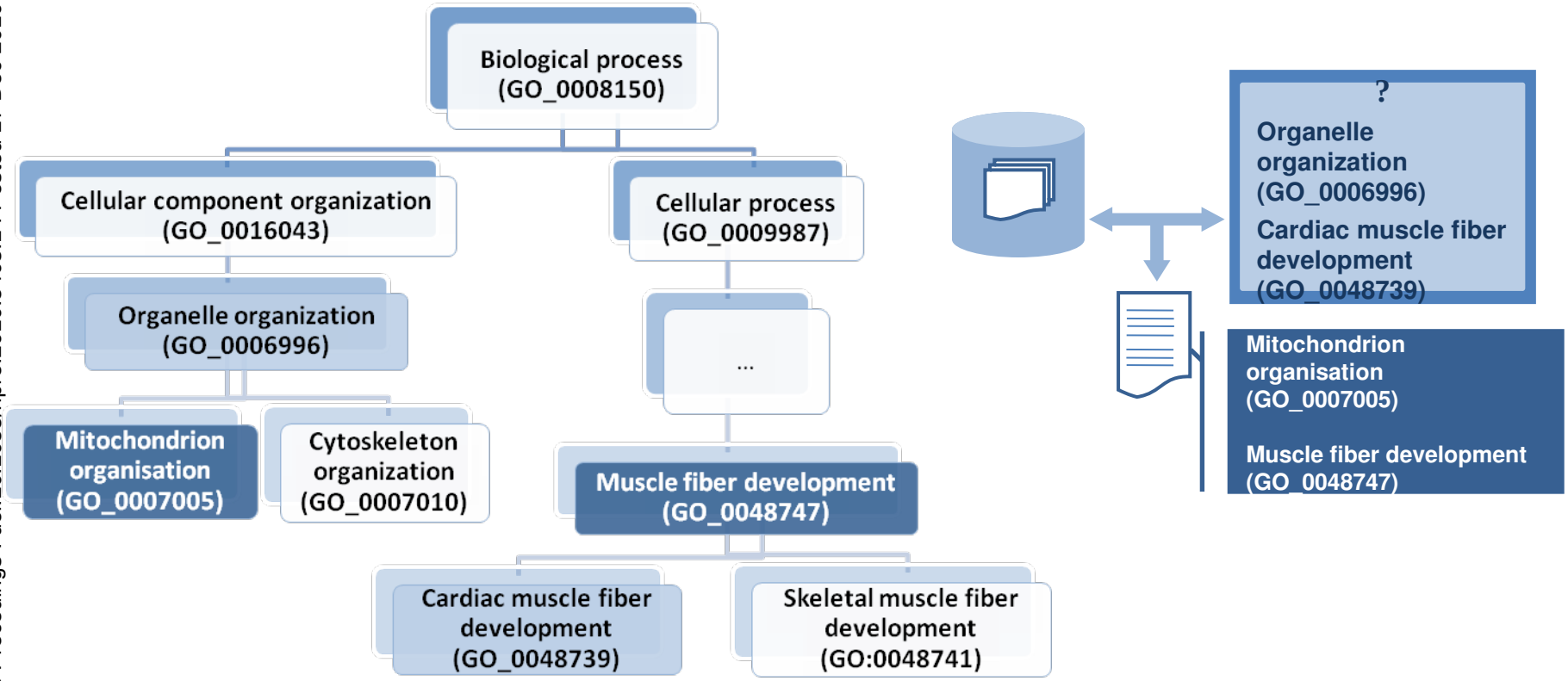
Relevance calculus between a document index and a query

- *Similarity between two concepts*
- *Relevance of a document with respect to a concept*
- *Relevance of a document with respect to a query*

Results visualization

Conclusion et perspectives

Hyponyms and hypernyms to avoid silences



- ⇨ Mix documents that match more or less the query
- ⇨ The selection may be difficult to understand

Context and objectives

Ontology based information retrieval

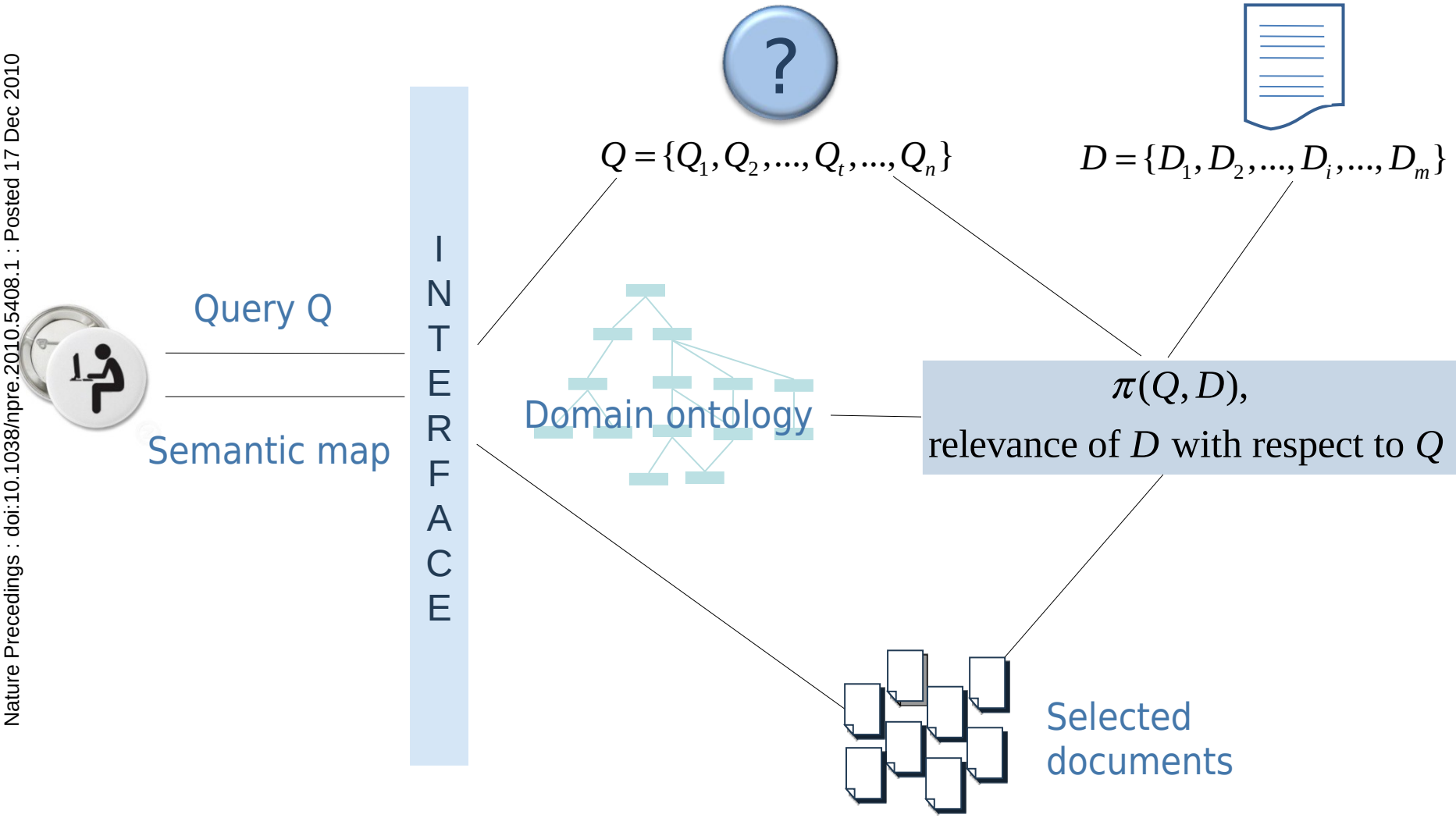
Relevance calculus between a document index and a query

- *Similarity between two concepts*
- *Relevance of a document with respect to a concept*
- *Relevance of a document with respect to a query*

Results visualization

Conclusion et perspectives

Relevance calculus between a document index and a query



Relevance calculus between a document index and a query

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$

⇨ Advantages

- Ranking of documents with respect to their relevance
- Detailed explanation of the document selection

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$

  - Several similarity measurements have been proposed in literature, this one is easy to understand (% of mutual hyponyms)

$$\pi_{JD}(C_1, C_2) = \begin{cases} \dfrac{hypo(C_1) \quad hypo(C_2)}{hypo(C_1) \quad hypo(C_2)}, & \text{if } C_1 \text{ } \square hypo(C_2) \text{ ou } C_2 \text{ } \square hypo(C_1) \\ 0, & \text{else} \end{cases}$$

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$

  - Best relevance between indexing concepts of document D and a query concept Q_t

$$\pi(Q_t, D) = \max_{0 \; i \; |D|} \pi(Q_t, D_i)$$

  - May be generalized by weighting the concepts D_i (using *evidence codes* in the Gene Ontology for example)

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$
  - Combine individual relevance scores to estimate an overall relevance of the document
    - Take user preferences into account: decision theory
    - Yager operator (with $q \in \mathbf{R}$)

$$Y_m(\pi(Q_1, D), .., \pi(Q_{|Q|}, D)) = \left(\frac{1}{|Q|} \sum_{t=1}^{|Q|} \pi(Q_t, D)^q\right)^{1/q}$$

q = 1, arithmetic mean,
q → 0, geometrical mean,
q = -1, harmonic mean,
q → +∞, max (OR generalization)
q → -∞, min (AND generalization)

Context and objectives

Ontology based information retrieval

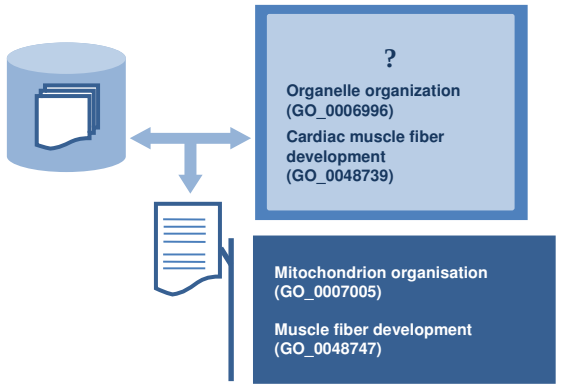
Relevance calculus between a document index and a query

- *Similarity between two concepts*
- *Relevance of a document with respect to a concept*
- *Relevance of a document with respect to a query*

Results visualization

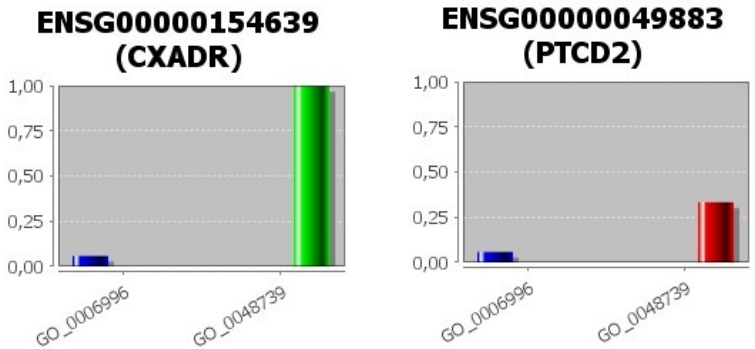
Conclusion et perspectives

- A document may be selected even if its index contains no terms of the query



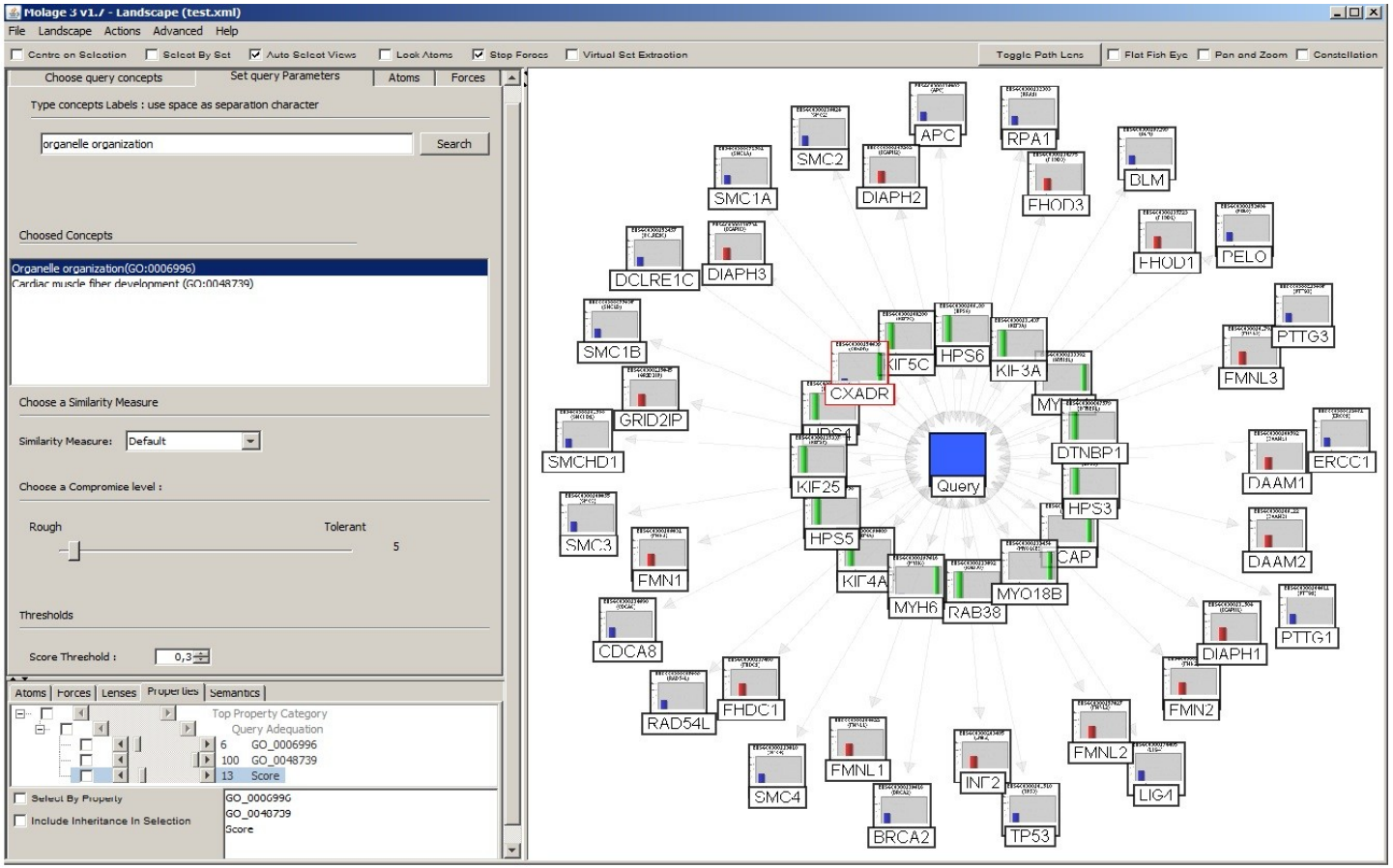
Explain the selection to the user: pictograms

- Each concept of the query is associated with a bar:
  - Its height is proportional to its relevance $\pi(Q_t, D)$
  - Its color says if Q_t
    - index the document (D)
    - specialize (is an hyponym of) D_i
    - generalize (is an hypernym of) D_i



Pictograms are displayed on a semantic map

- Their physical distance to the query is proportional to their relevance score: $\pi(Q, D)$
- Visualization and navigation: fit the human cognitive limits (lens, number of results, relevance threshold…) and help the user (selection of concept for the query…)



Context and objectives

Ontology based information retrieval

Relevance calculus between a document index and a query

- *Similarity between two concepts*
- *Relevance of a document with respect to a concept*
- *Relevance of a document with respect to a query*

Results visualization

Conclusion et perspectives

Results

- Find more documents (avoid silences)
- Improve the relevance: documents ranking
- Explain relevance calculus (diagnose)
- Visualize overall results
- Interaction with the list of retrieved documents: customize user preferences
  - ⇨ Iterative improvement of the query

Perspectives

- Improve CHI
- Suggest query reformulation
  - From documents selection by the user (weighting + complement)
  - Underline query terms that are discriminated
- Test several semantic distance calculus on different benchmarks (TREC, Much more…)
- Improve visualization
  - Filter the displayed results using sub-ontologies extraction
  - Propose a view of the results underlining clusters
- Propose an online version

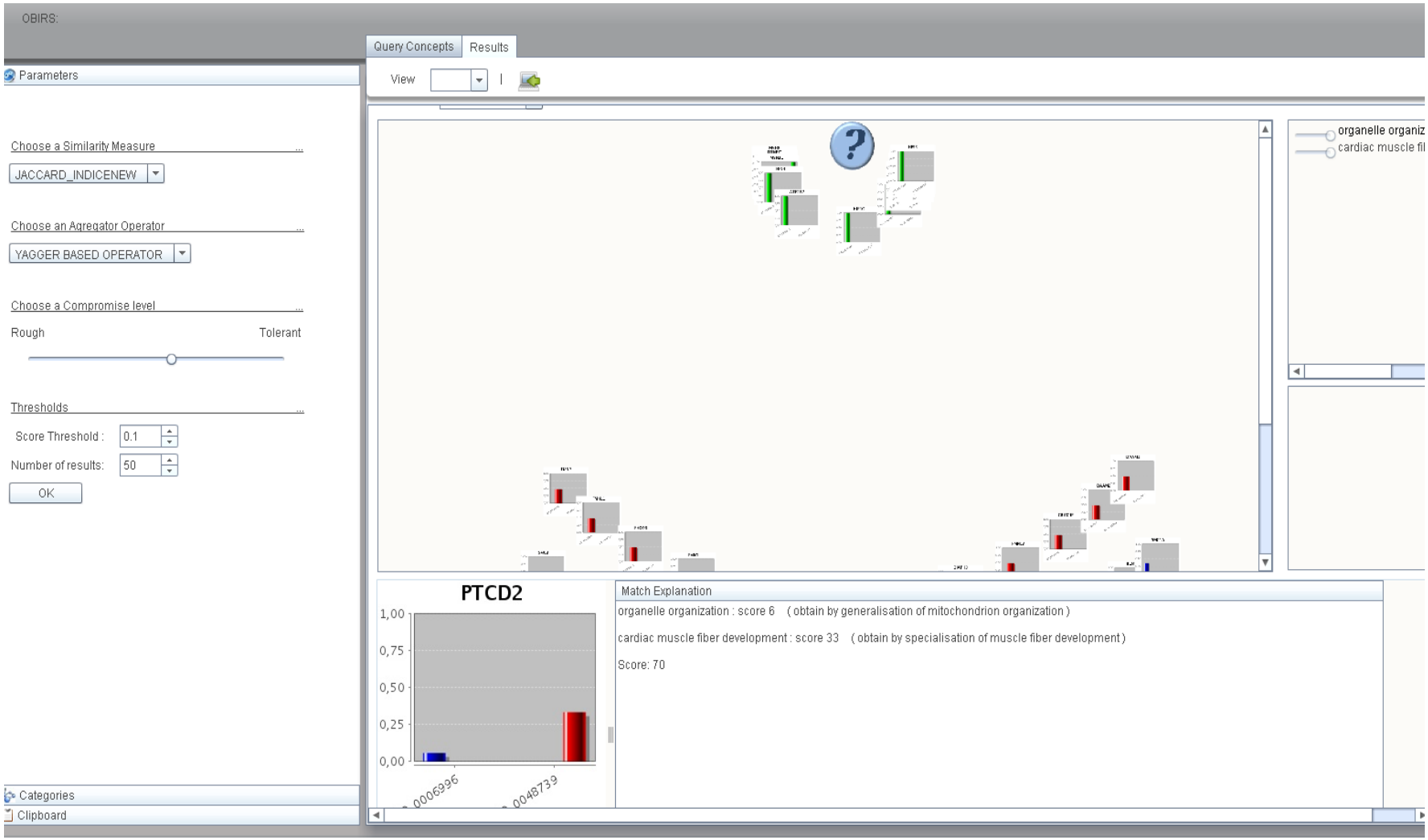
User Centered and Ontology Based Information Retrieval System for Life Sciences

sylvie.ranwez@mines-ales.fr
vincent.ranwez@univ-montp2.fr
mohameth.sy@mines-ales.fr
jacky.montmain@mines-ales.fr
michel.crampes@mines-ales.fr

ÉCOLEDESMINESD'ALÈS
ALÈS•NÎMES•PAU

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

OBIRS on line: http://www.ontotoolkit.mines-ales.fr/ObirsClient/



Nature Precedings : doi:10.1038/npre.2010.5408.1 : Posted 17 Dec 2010