CominWeb LookinLabs

William Dedzoe

Albert Benveniste

Bertrand Le Marouille

Mauricio Urraco

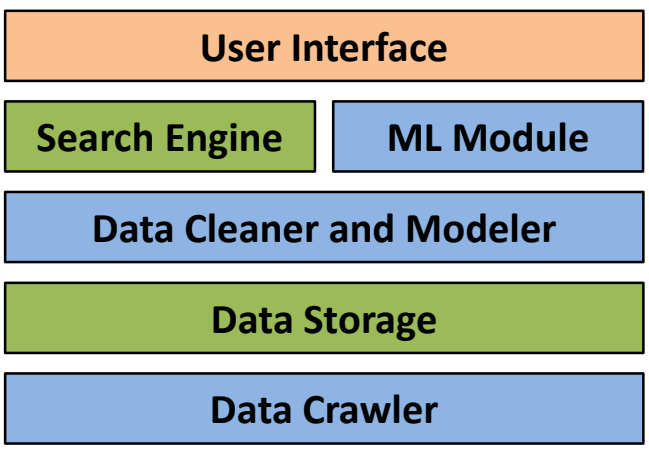
Jean Hany

Inria

Motivating Context: CominLabs

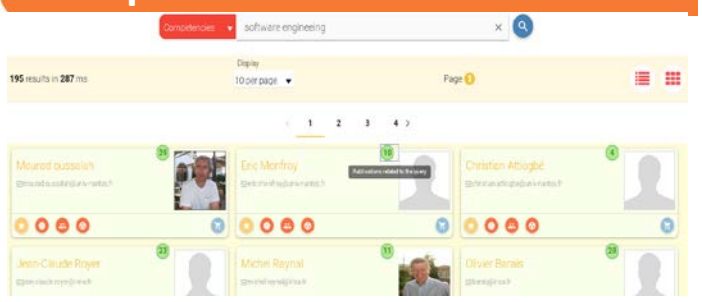
- **Approximately 500 researchers, among them 350 are registered in the database through their bibliographical links**
- **Objectives**
  - Scientific search engine
  - Competencies and profiling of researchers
  - Links to bibliography
  - "Zero" maintenance
- **Approach**
  - Machine learning and big data technologies
- **Data: bibliographical data base**
  - Title + Abstract of publications

Lookinlabs main components



- **Data crawler**
  - crawls publications from open publication databases
- **Data storage**
  - **MySQL** database storing all the publications crawled by the data crawler
- **Data cleaner and modeler**
  - cleans and enriches data and then puts the clean data into **Couchbase** database for indexation
- **Sear engine**
  - **ElasticSearch** engine indexes data coming from the data cleaner and modeler layer
- **Machine Learning (ML) module**
  - Contains our algorithms for text mining, e.g., topics detection (this module is based on **nltk**, **gensim** and **scikit-learn**)
- **Graphical User Interface**
  - built using **Angular Material**

Competencies search

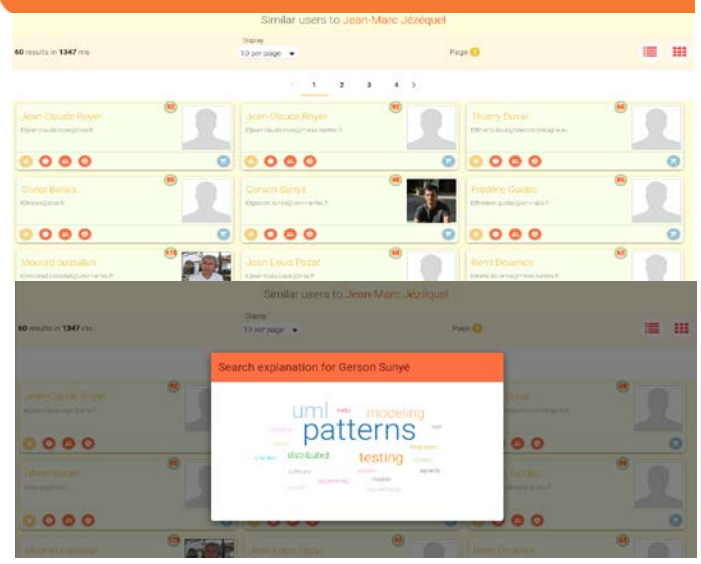


Searching for individuals working on a subject described via free keywords (no predefined ontology)

Automatic profile generation

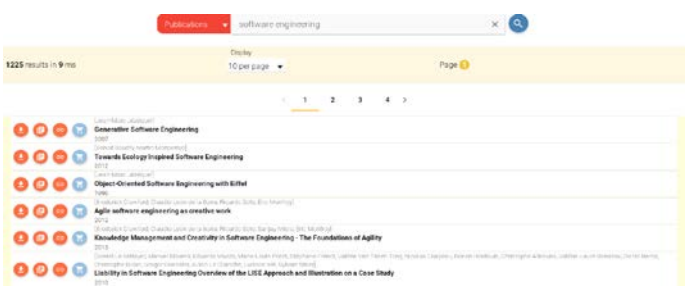
Automatic identification of the rich topics addressed

Similar researchers



Identification of similar researchers and explanation of why a given researcher was found similar

Search publications



In progress: supporting HAL

- **What is HAL**
  - Scientific An open common archive platform of French research institutes and universities
  - 1 197 186 publications (62% have an abstract)
- **What do we do for HAL?**
  - Data cleaning and data enrichment
  - Data modeling
  - Data indexation
  - Data mining
- **Synergies with other developments for HAL**
  - Grobid:
    - machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI-encoded documents with a focus on technical and scientific publications
  - AnHALytics:
    - publication centric search engine operating on full pdf HAL documents, daily updated
    - uses Grobid to structure the pdf
    - desambiguation using Wikipedia
- **ScanR**
  - Official registry of all research institutions, both public or private (25000 registered); developed and operated by the government
  - Gathers information from web sites, patents, and publications, for each institution
  - ScanR comes with a search engine

Future work regarding HAL

- **Query individuals & labs**
- **New difficulties**
  - Labs possess multiple names (acronyms, expanded, variations)
  - Labs evolve in time with ancestors and children
  - Topics are wider in scope (tune the algorithms differently)
- **Schedule: march 2017**

References

https://**scanr**.enseignementsup-recherche.gouv.fr/

http://traces1.saclay.inria.fr/**anHALytics**/search/

https://**grobid**.readthedocs.io/en/latest/

https://www.**elastic**.co/fr/

http://www.**couchbase**.com/
http://www-fr.**mysql**.com/
http://www.**nltk**.org/
https://radimrehurek.com/**gensim**/
http://**scikit-learn.org**/
https://**material.angularjs**.org/latest/

http://www.cominlabs.ueb.eu/fr/LookinLabs