DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Visualització i Gestió de la Informació Científica

www.dama.upc.edu

www.sparsity-technologies.com

*sparsity technologies
performance in action

Josep L. Larriba Pey, larri@ac.upc.edu

DAMA-UPC: research universitary group

Our goal

Research and Technology Transfer on Managing Very Large Data Volumes

Since 1999

Founded inside UPC

The beginnings

22 researchers and developers

The team

Funding, agreements and collaborators

ICO Institut Català d'Oncologia · ca technologies · IBM · HAVAS MEDIA · acceso · ancert Agencia Notarial de Certificación · bmat · Gobierno de España · Ministerio de Ciencia e Innovación

Our software

IBM Faculty Awards (2004, 2009, 2012)

IBM PhD Award (2004)

CINC prize for novel enterpreneurs (2009)

Bdigital prize for Universities, 2011

Awards

> 75 research papers 6n patents

Publications

JGIC 2012

Visualització i Gestió de la Informació

1. Fonts de dades
2. Integració/qualitat de la informació
3. Consultes de valor afegit
4. Automatització de processos
5. Visualització

1. Fonts de dades

- Bibliogràfiques:
  - Obertes: PubMed, ArXiv, DBLP
  - Propietàries: Scopus, Thomson-Reuters, ACM, IEEE, etc
- Patents:
  - Obertes però pagant
- Xarxes socials:
  - APIs per accedir a la informació social
  - Relació entre persones
- Altra informació:
  - Cordis
  - Proteïnes
  - Productes químics

2. Integració/qualitat

- Qualitat de les fonts de dades
  - Errors tipogràfics
  - Falta d'informació
- Desambiguació d'autors
  - Gran problema actual
- Citacions incomplertes
  - Molt car de mantenir (Scopus >20K artícles diaris)
- Preservar la privacitat de les dades donant accés a les metadades

- Solucions possibles
  - Us de tecnologia de grafs i desambiguació
  - Metaintegradors de dades que millorin les dades a partir de la interacció dels usuaris

3. Consultes de valor afegit

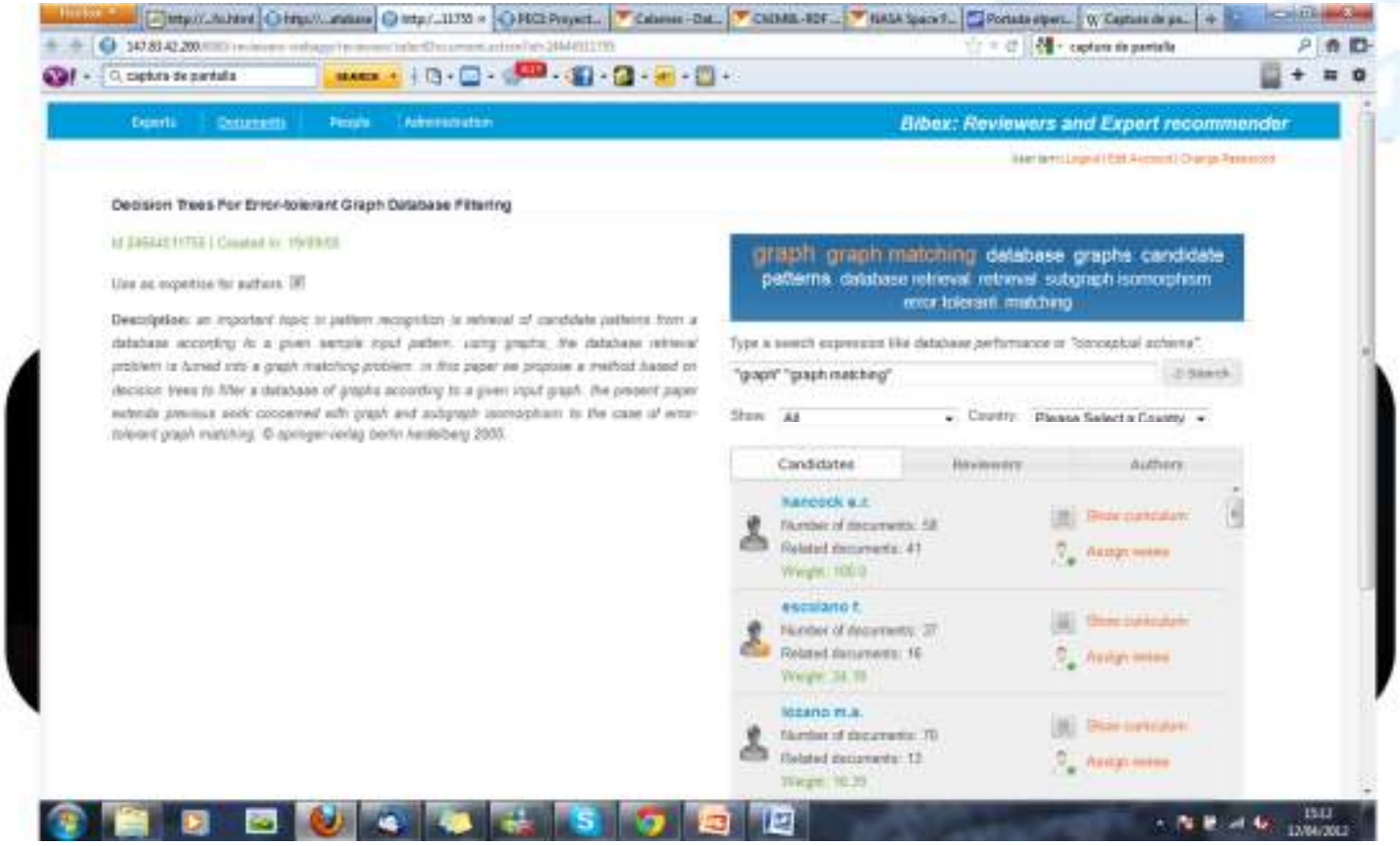
Amb l'us de paraules clau, resums o metadades en grl.

- Reputació
- Cerca d'obres més valuoses (citacions, reputació dels autors)
- Evolució històrica d'àrees científiques a partir de paraules clau

- Solucions possibles:
  - Us de tecnologia de gestió de grafs
  - Consultes avançades

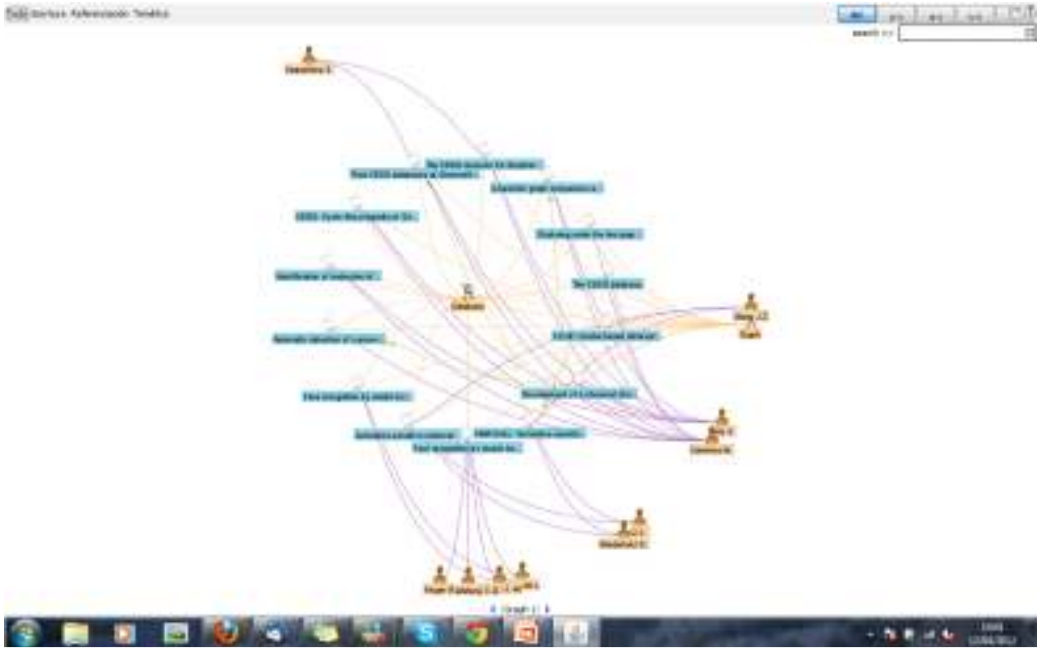
4. Automatització de processos

- Cerca i gestió de revisors



5. Visualització

- Com fer obvis els problemes causats per les dades de forma visual
- Com mostrar diferents fonts de dades
  - Aportant valor afegit
  - Mostrant que les dades són de fonts diferents
  - Integrant sense molestar als propietaris



Solucions al problema

- DAMA-UPC i Sparsity proporcionen tecnologia:
  - Gestió de grafs, Dex
  - Reputació
  - Cerca de revisors
  - Suport a l'elaboració de projectes, Sciencea

Altres preguntes

- Donat que les dades amb una mínima qualitat són cares (Scopus, Thomson...), quina qualitat es pot aconseguir amb dades públiques (DBLP, PubMed, etc)?
- Podem aconseguir mètodes alternatius per a donar qualitat a les dades publiques que siguin ràpids, barats (sense intervenció humana) i ens donin fiabilitat?
- Què volem saber de la informació científica? Com volem que es visualitzi, en forma de llistat o en forma gràfica?

Altres preguntes (cont)

- Integrar informació de forma ràpida és possible, i.e. Scopus rep més de 20K artícles per dia i els ha d'integrar a la seva base de dades de milions d'articles?
- Que ens aportaria tenir bases de dades integrades amb fons bibliogràfiques, de patents, d'institucions, d'experiments científics, etc?
- Com podriem representar informació integrada de forma visual?

Any questions?

Contact e-mail: larri@ac.upc.edu

DAMA Group Web Site: http://www.dama.upc.edu

JCC 2011

TECNIO
Be tech. Be competitive

DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA