SEARCHY: A Metasearch Engine for Heterogeneous Sources in Distributed Environments

David F. Barrero
M. Dolores Moreno
Óscar García Población
Ángel Moreno

DC 2005
Madrid, Spain
14th September, 2005

Roadmap

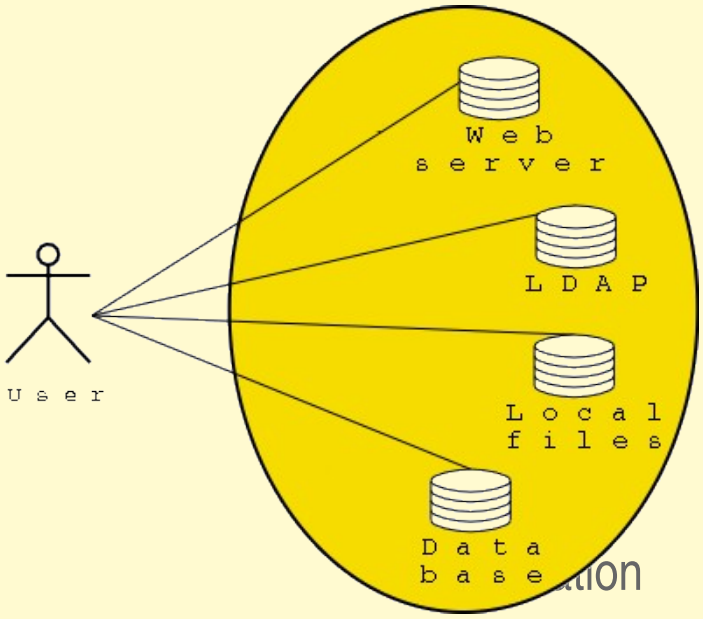
- Introduction
- Searchy, a metasearch engine based on Dublin Core
  - What's Searchy for?
  - How Searchy works?
  - A simple example
  - A more complex example
- Study case: RedIRIS
- Work in progress
- Conclusions

Introduction (I)

- Documents are a key value for success in any organisation
  - There are many solutions that help in this task
  - Each solution uses to be suitable for a context
- Many times, using a single document management system is not possible
  - Documents might be associated with legacy services
  - Documents nature may require to use different information systems
- In this situation homogeneity is lost
  - Documents format may be different
  - Metainformation associated with the document may have the same problem
    - A .doc file has not the same metainformation like a PDF document

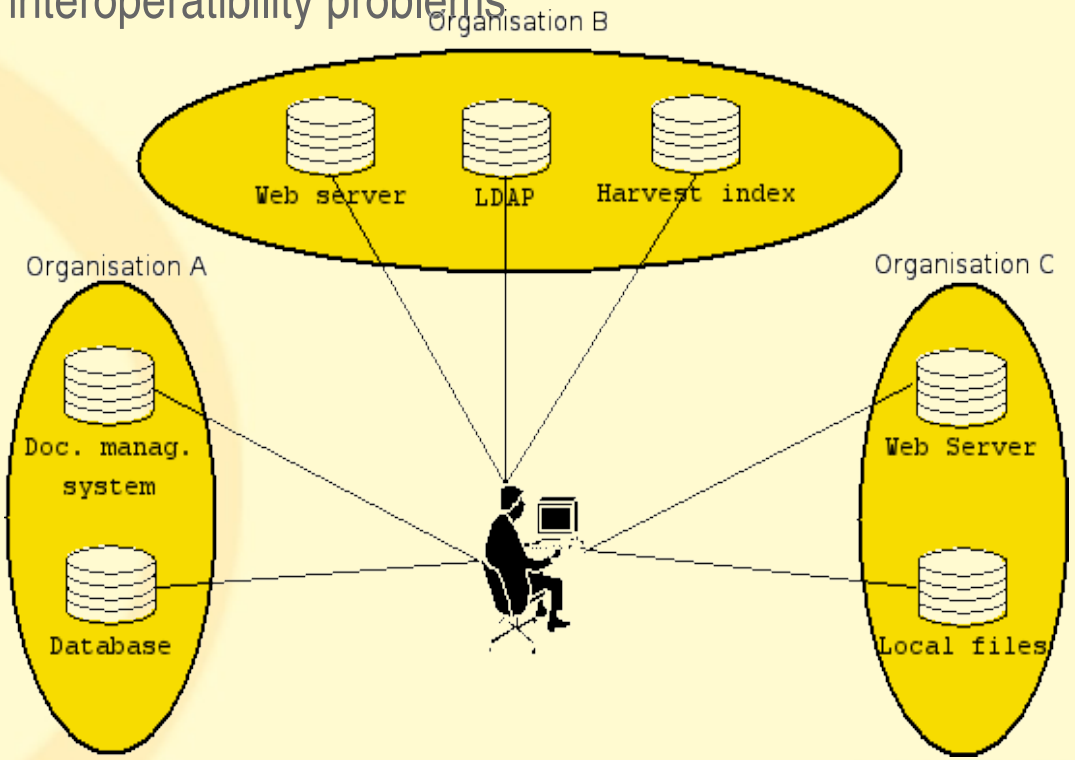
Introduction (II)

- Heterogeneity in information systems may be a major problem
  - **Expensive middleware** to grant interoperatibility
  - **Increased complexity** to locate documents
  - **Different interfaces** to access documents
  - **Several simple** user interfaces or one
- ... however sometimes we need heterogeneous systems with an uniform access
  - Our proposal to solve this problem of information systems



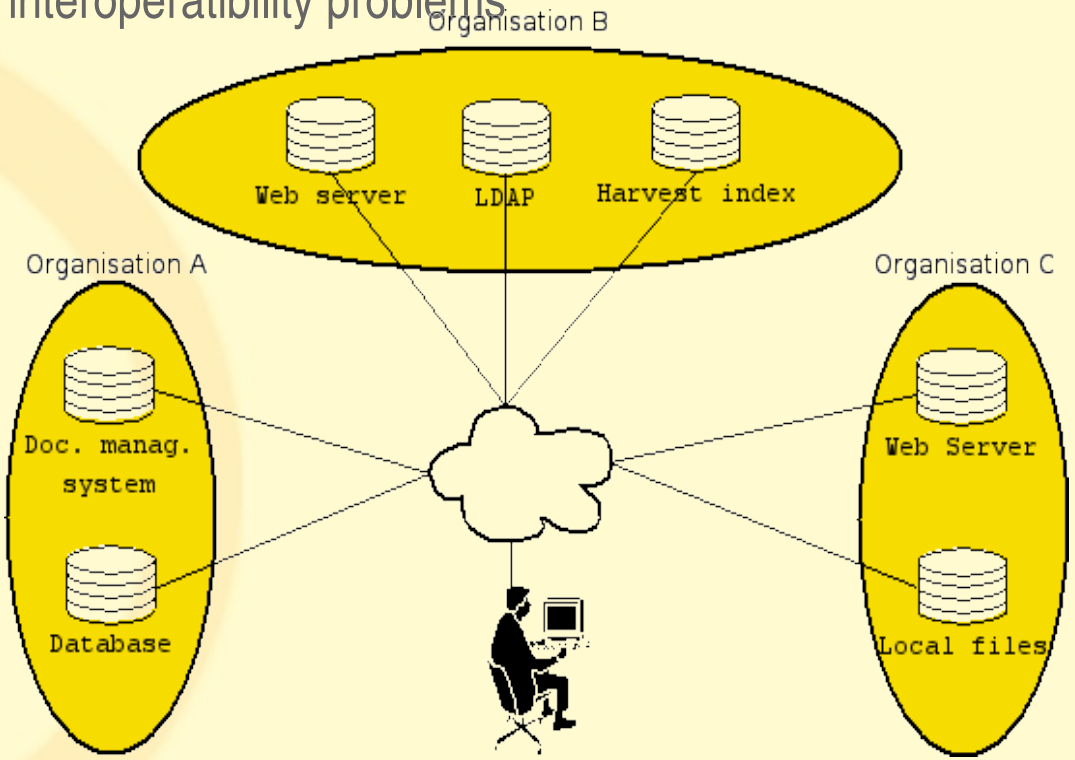
Introduction (III)

- The problem is more complex when there are involved several organisations
  - Increased interoperatibility problems



Introduction (III)

- The problem is more complex when there are involved several organisations
  - Increased interoperatibility problems



Searchy: a distributed metasearch engine

What's Searchy for? (I)

- It is a project leadered by the University of Alcalá and RedIRIS (the Spanish National Research and Educational Network)
- Searchy is a distributed middleware that integrates information semantically
  - Actually, Searchy may be seen like a metasearch engine, i.e, a search engine that searches across different search engines
- Searchy performs several tasks, it ....
  - **translates** a query into a local format,
  - **extracts** local metainformation,
  - **maps** local metainformation format into Dublin Core,
  - and finally **integrates** the results.

Searchy: a distributed metasearch engine

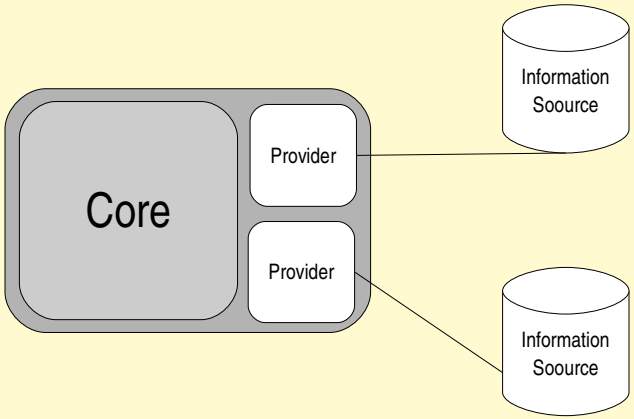
What's Searchy for? (II)

- In short, there may be different document management systems, but Searchy users will perceive them like one monolitic system, with one metadata model
  - Dublin Core is the shared metainformation model
  - Dublin Core provides a great flexibility in the type of document that may be described with Searchy
  - The mapping must be done manually by Searchy administrator
- From the user's point of view, Searchy is only a search engine, but ... who is the user?
  - Searchy is not suppossed to be used directly by end users
  - It is just a middleware software

Searchy: a distributed metasearch engine

How Searchy works?

- La basic unit is the agent, which is composed by...
  - a *core*, contains the common funcionalities
  - a *provider*, it access information sources
  - one or more *information sources*, it is the backend
- Information sources supported:
  - SQL data bases, LDAP directories, Harvest brokers and the Google API
- Providers decouple information sources from the core
  - This approach is highly flexible: Searchy may be used with almost all sort of information and backends
  - If you can read information, you can integrate it with Searchy



Searchy: a distributed metasearch engine

A simple example

- Agent that integrates a relational database and a directory

Relational database table

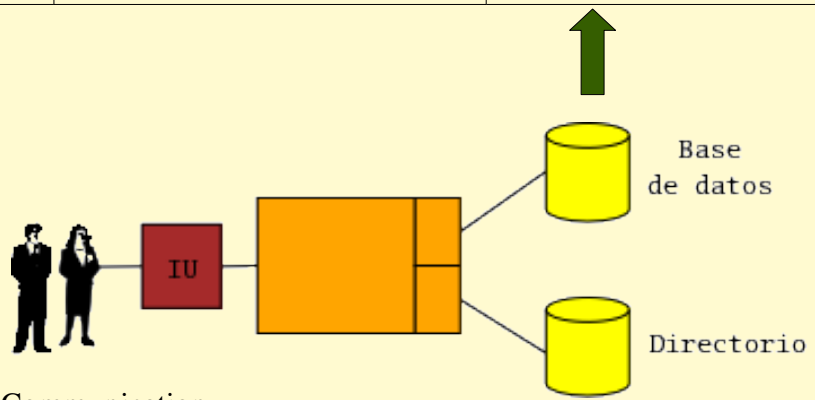
AUTHOR	*BOOK*	*KEYWORDS*
Homer Simpson	Memories of a donuts eater	Sprinfield, Simpson family, donuts
Darth Vader	The Dark Side for dummies	The Force, The Empire, sith lords

DC representation

DC.Title = Memories of a donuts eater
DC.Creator = Homer Simpson
DC.Subject = Sprinfield, Simpson family, donuts

DC.Title = The Dark Side for dummies
DC.Creator = Darth Vader
DC.Subject = The Force, The Empire, sith lord

DC.Title = Improving Rasterization Using Psychoacoustic Communication
DC.Creator = John S. Random
DC.Date = The Force, The Empire, sith lord
DC.Publisher =International Journal of Something



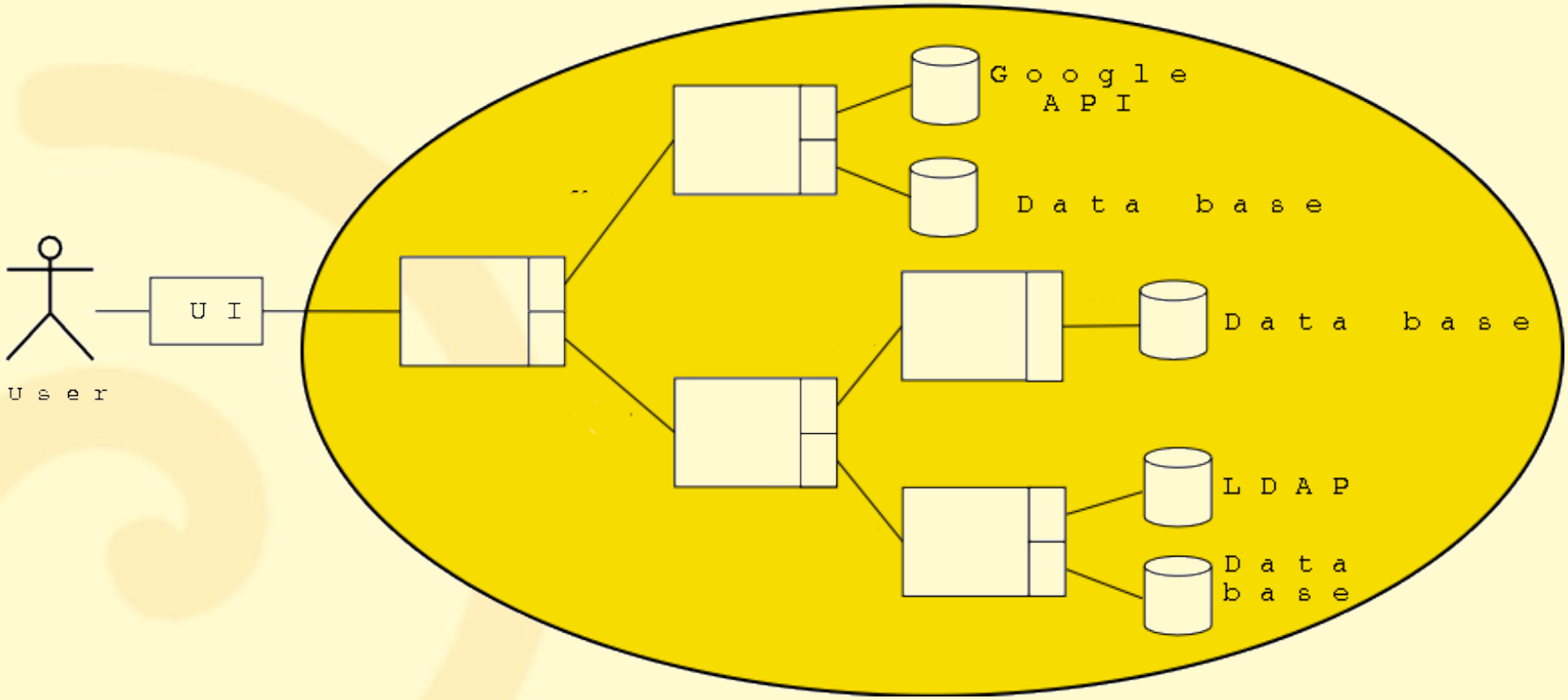
Directory tree

dc=es
dc=example
uid=0001
title=Improving Rasterization Using Psychoacoustic Communication
author=John S. Random
journal=International Journal of Something
creationDate=1-06-77

Searchy: a distributed metasearch engine

A more complex example

- The beauty is in the fact that agents may communicate with other agents



Case study

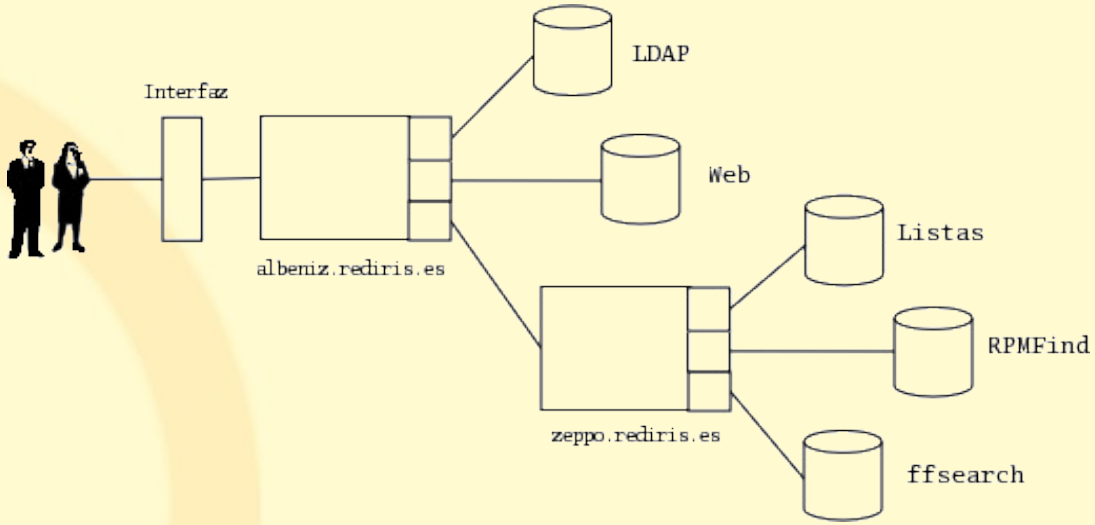
Systems integration in RedIRIS (I)

- Generic resource approach instread of a document centric approach
- There is a wide range of resources
  - Web, FTP, mailing lists, LDAP directories...
  - Multiple search engines available in our web
    - An unified access for searching resources looks a better solution
- The environment is quite complex
  - Diversity of hardware and software
  - Heterogeneuos information systems
  - Strict security policies
  - High number of users

Case study

Systems integration in RedIRIS (II)

- Solution:
  - Federation of document search engines using Searchy



- It is in a beta stage
- Available in http://www.rediris.es/busquedas/searchy/search.en.phtml

Work in progress

- Improve the scalability
  - Using alternative transport mechanisms like P2P or multicast
- Use intelligent agents
  - Learn user profiles by using case based learning
  - Ordering of resources
  - Improve quering semantics
- Increase the number of information sources supported
  - Google desktop, Beagle, ...
- Access control management
- Improve the user interface

Conclusions

- Two great advantages
  - Searchy respects legacy systems and autonomy
  - Almost all types of information may be supported
- Not all advantages in Searchy...
  - It is not as lightweight as using each information system alone
  - Scalability limited in this moment by the nature of communications
- The ideal situation in which Searchy might be used is
  - Several independent organisations involved
  - Each organization has legacy heterogeneus information systems

Thanks for your attention!

:)