

MUMIA: INTEGRATING IR TECHNOLOGIES FOR PROFESSIONAL SEARCH

Mike Salampasis

Marie Curie Fellow

Vienna University of Technology

Institute of Software and Interactive Systems

ESSIR 2013

Outline



- **MUMIA**
- Professional Search: Introduction and Some Terminology
- Integrated Search Systems
- A General Framework for Integrated Professional Search Systems
- Case Study – Putting things to work
- Open Problems

Scientific context and objectives

- The aim of the Action is to coordinate and support the interaction and harmonization of high quality research at a European level in the field of multilingual and multifaceted interactive information access with a view to contribute to the development of next-generation (professional) search systems.
- Influence the R&D of leading state-of-the-art projects related to professional search
- Patent search is used as unifying testbed

MUMIA Working Groups

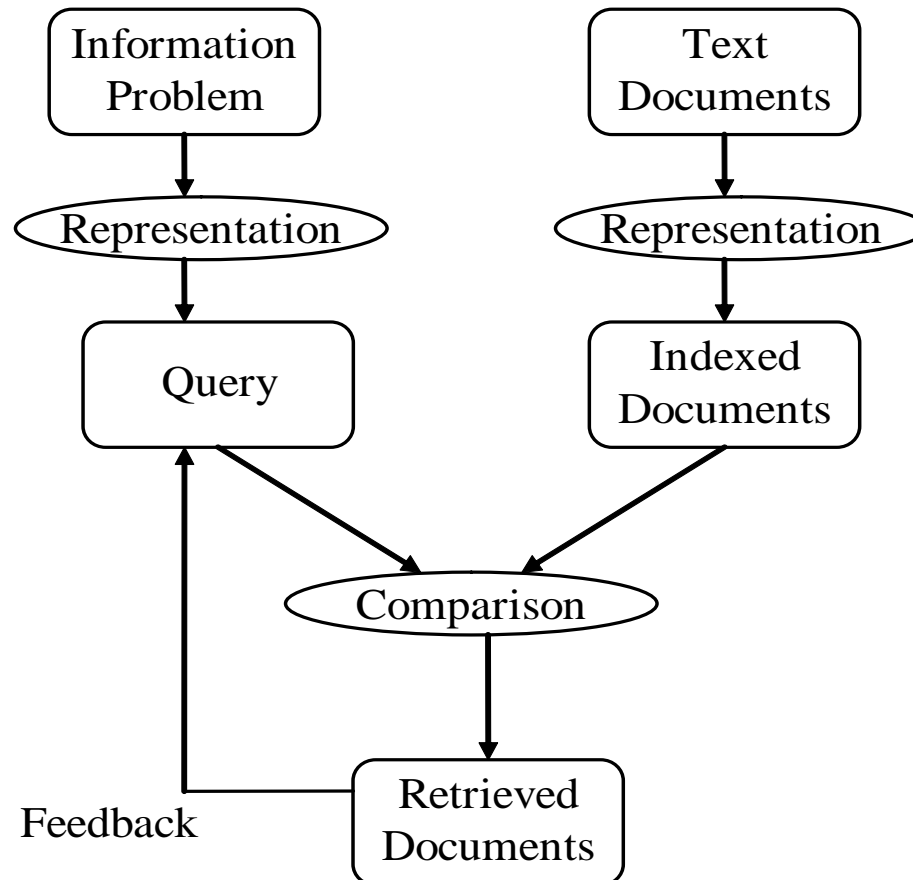
- ❑ WG1: Integrating and Managing Language Resources.
- ❑ WG2: Processing Infrastructures for IR and MT.
- ❑ WG3: User Centred Aspects of MUMIA.
- ❑ WG4: Semantic Search and Faceted Search, Visualization.
- ❑ WG5: Distributed and Social Search.

Outline



- MUMIA
- **Introduction to Professional Search and Some Terminology**
- Integrated Search Systems
- A General Framework for Integrated Professional Search Systems
- Case Study – Putting things to work
- Open Problems

Basic Information Retrieval Processes



The Classic Model for IR, augmented for the web (Andrei Broder, 2003)

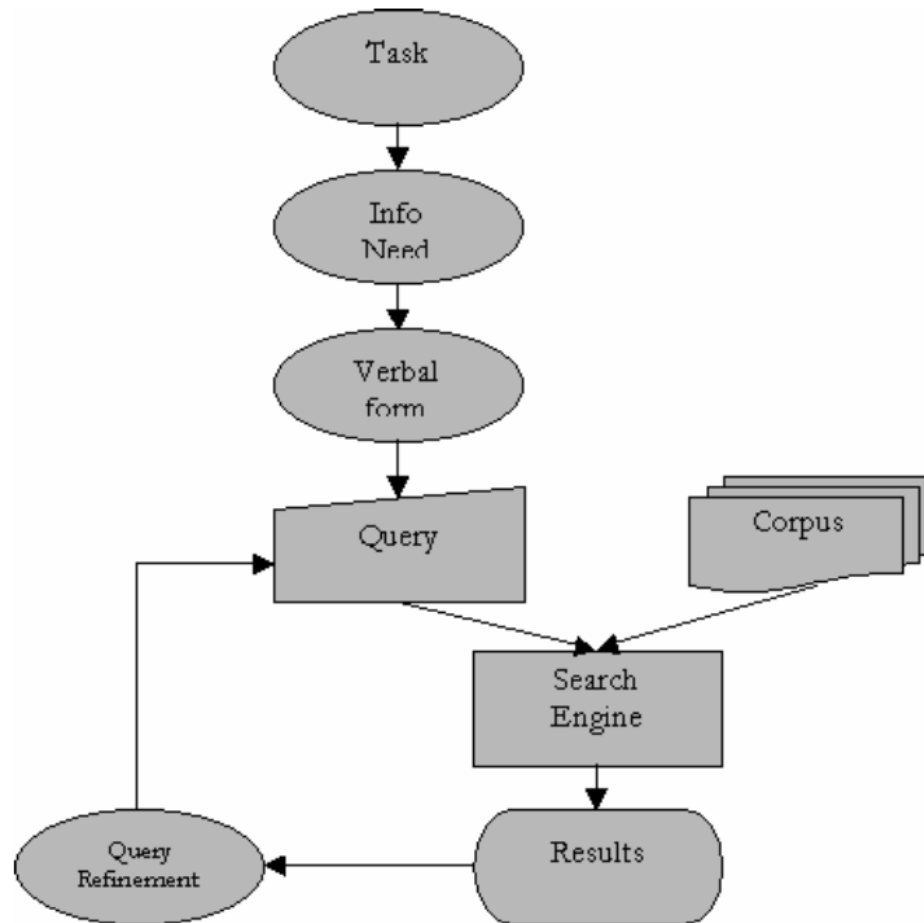
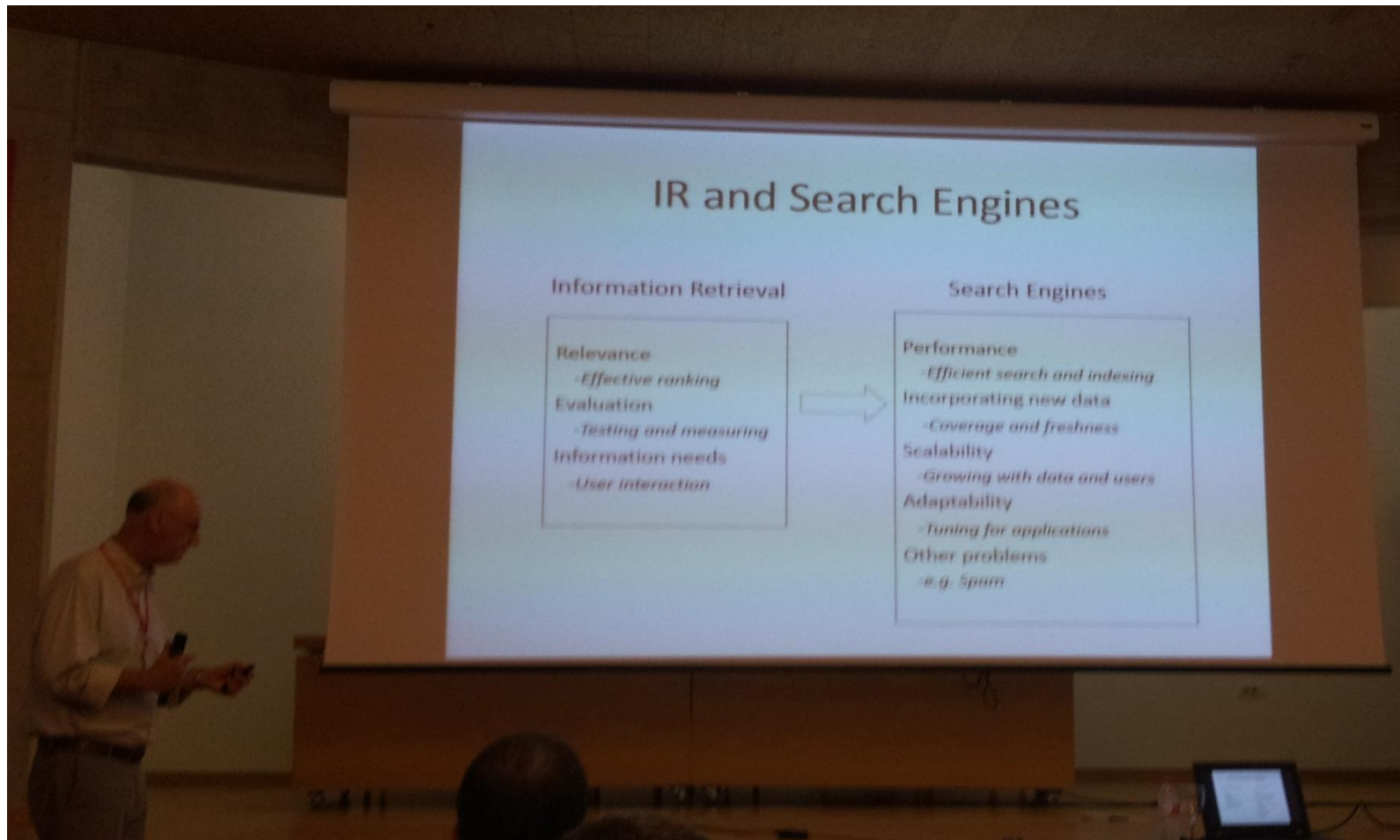


Figure 2. The classic model for IR, augmented for the web.

From IR to Search Engines to ...



From Croft's talk this morning

Professional Search

- Professional Search is search in the workplace or search for a professional reason or aim and can occur in many different domains (e.g. patent, medical, engineering, scientific literature search, media reports)
- There are a number of important parameters and characteristics that differentiate professional search from web search

Status of Professional Search

- Search technologies are used for **professional search** for more than 40 years as an important method for information access
- Despite the tremendous success of web search technologies, there is a significant skepticism from professional searchers and a very conservative attitude towards adopting search methods, tools and technologies beyond the ones which dominate their domain.
- An example is patent search where professional search experts typically use the Boolean search syntax and quite complex intellectual classification schemes

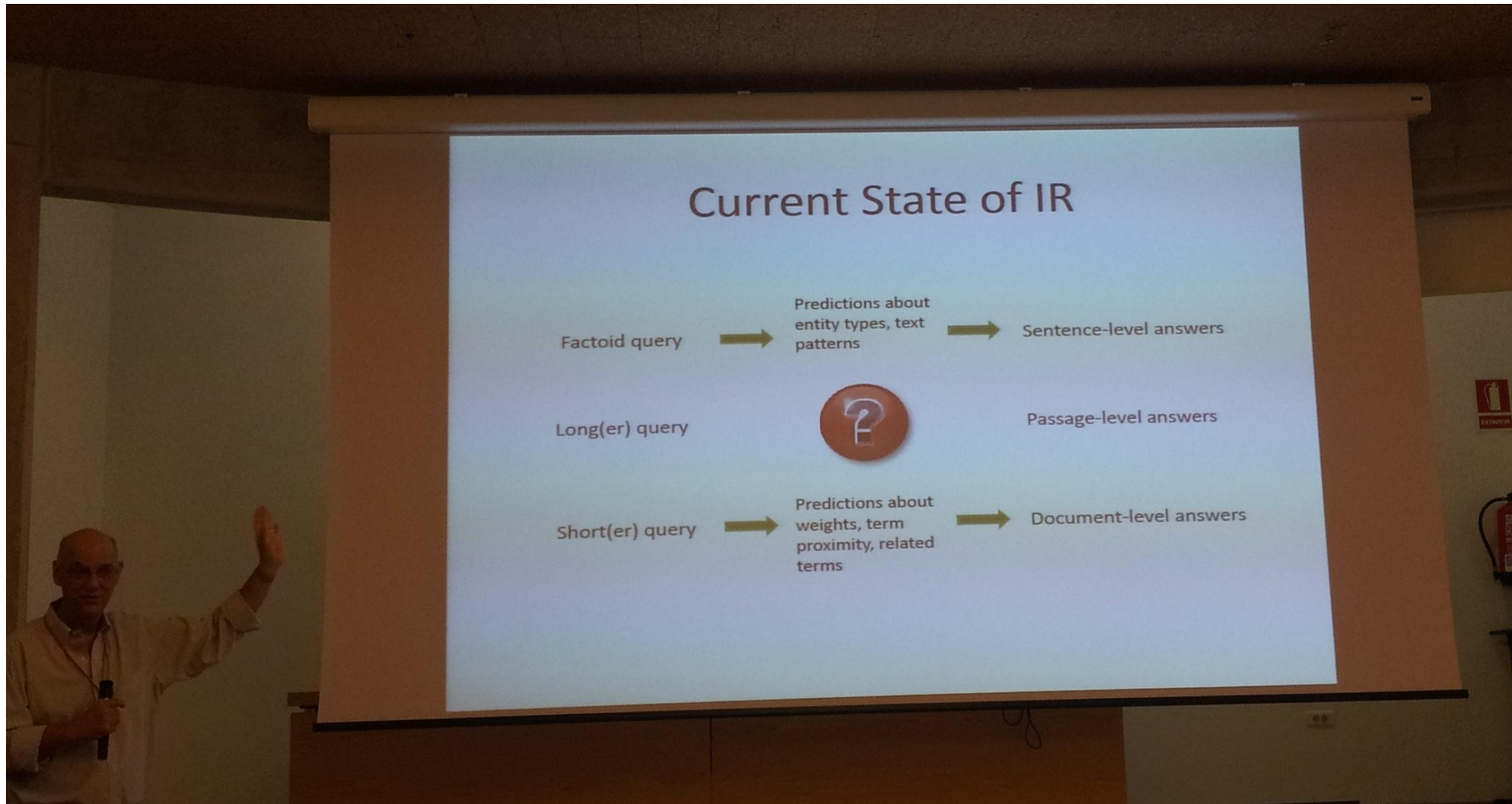
Professional Search vs. Web Search

- lengthy search sessions (even days) which may be suspended and resumed,
- the notion of relevance can be different,
- many different sources will be searched separately, and
- focus is on specific domain knowledge in contrast to public search engines which are not focused on expert knowledge.

Professional Search vs. Web Search

- Often high recall is important.
- Reason about how the results have been produced
- Reproducibility of a search process (e.g. patent searcher is required to prove the sufficiency of the search in court at a later stage).
- Classification schemes and metadata are heavily used because it is widely recognized that once the work of assigning patent documents into classification schemes is done, the search can be more efficient and language independent.

Short, Long, Factoid Queries and...



Information Needs of Professionals

- How Much Is My Patent Worth If I Sell It ?
- Shall my company invest 10 million EUR in plastic packaging business ?
- My company wants to develop a new coffee machine. Which are the technical areas related to the development of such apparatus? I want to know the prior art of the last 10 years.

There are many variables the patent professional has typically to work with

- For example in a typical patent search
 - Technical subject: A + B + C ...
 - Databases
 - Keywords
 - Codes: e.g. IPC, ECLA, Manual Codes
 - Patent assignee and inventor names
 - Patent countries, date ranges

There are many variables the information professional has to work with

➤ An even more

- OR
- AND (are you sure?)
- OR and AND (collections and intersections)
- Proximity operators
- NOT (be careful !)
- Forward and backward citation searches

Outline



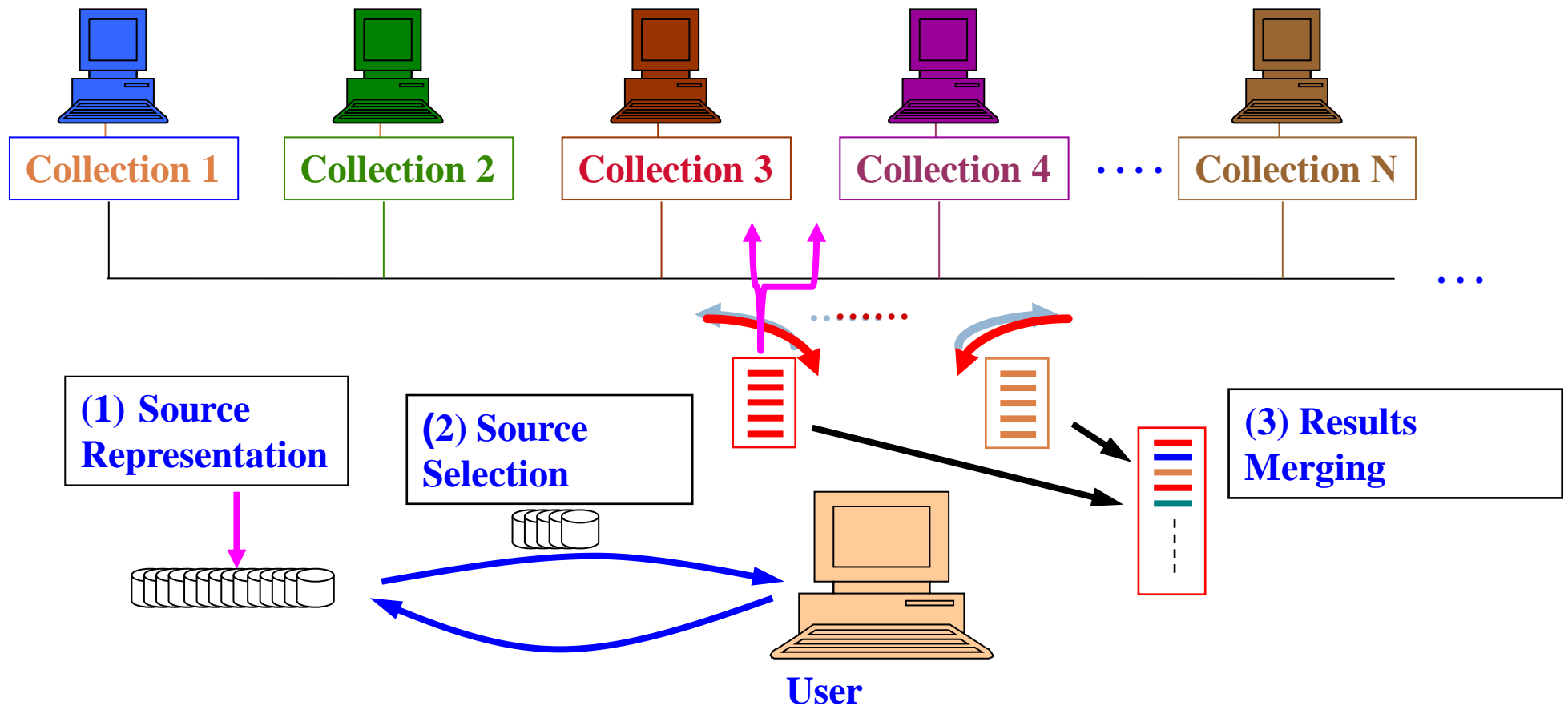
- MUMIA
- Introduction to Professional Search and Some Terminology
- **Integrated Search Systems**
- A General Framework for Integrated Professional Search Systems
- Case Study – Putting things to work
- Open Problems

Some terminology clarification

- Federated search
- Aggregated search
- Integrated search

Federated Search - Distributed IR

Elements composing a Distributed Information Retrieval System



Motivation for federated search



- Search Hidden/Deep web collections
 - ▣ Collections not (easily) crawlable
- Access up-to-date information and data
- In theory it can be more scalable than centralized approaches
- It can also be more effective (cluster hypothesis)

Aggregated search

- Federated approach for the web
 - ▣ Meta-search engine combines the results of different search engines into a single result list
 - ▣ Vertical search – also known as **aggregated search** – add the top-ranked results from relevant verticals (e.g. images, news, videos, maps, structured information) to typical web search results

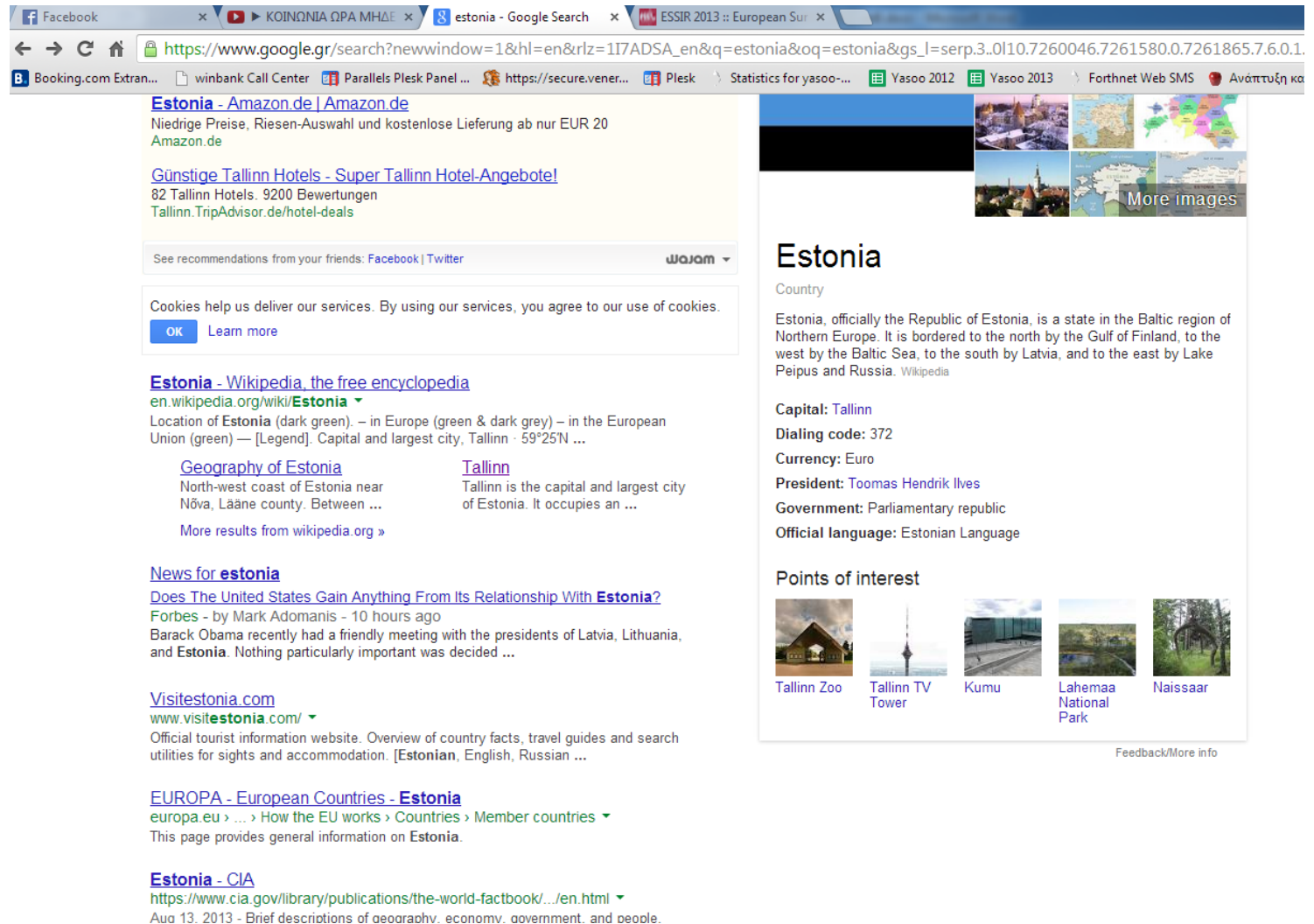
Aggregated Search

An example of aggregated search using the term

Estonia

You can get:

- Home page
- Video
- Wikipedia
- Structural info
- Images
- news



The screenshot shows a web browser with multiple tabs open: Facebook, KOINONIA OPA MHΔE, estonia - Google Search, and ESSIR 2013. The main content area displays search results for "estonia". The results are organized into a grid-like format with various links and images.

Search Results:

- Estonia - Amazon.de | Amazon.de**
Niedrige Preise, Riesen-Auswahl und kostenlose Lieferung ab nur EUR 20
Amazon.de
- Günstige Tallinn Hotels - Super Tallinn Hotel-Angebote!**
82 Tallinn Hotels. 9200 Bewertungen
Tallinn.TripAdvisor.de/hotel-deals
- Estonia - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Estonia
Location of **Estonia** (dark green). – in Europe (green & dark grey) – in the European Union (green) — [Legend]. Capital and largest city, Tallinn · 59°25'N ...
- Geography of Estonia**
North-west coast of Estonia near Nõva, Lääne county. Between ...
More results from wikipedia.org »
- Tallinn**
Tallinn is the capital and largest city of Estonia. It occupies an ...
- News for estonia**
Does The United States Gain Anything From Its Relationship With Estonia?
Forbes - by Mark Adomanis - 10 hours ago
Barack Obama recently had a friendly meeting with the presidents of Latvia, Lithuania, and Estonia. Nothing particularly important was decided ...
- Visitestonia.com**
www.visitestonia.com/
Official tourist information website. Overview of country facts, travel guides and search utilities for sights and accommodation. [Estonian, English, Russian ...
- EUROPA - European Countries - Estonia**
europa.eu > ... > How the EU works > Countries > Member countries
This page provides general information on Estonia.
- Estonia - CIA**
https://www.cia.gov/library/publications/the-world-factbook/.../en.html
Aug 13, 2013 - Brief descriptions of geography, economy, government, and people.

Images:

- Tallinn Zoo
- Tallinn TV Tower
- Kumu
- Lahemaa National Park
- Naissaar

Feedback/More info

What is integrated search?

Many definitions usually centered around the idea of a single point of search for multiple sources

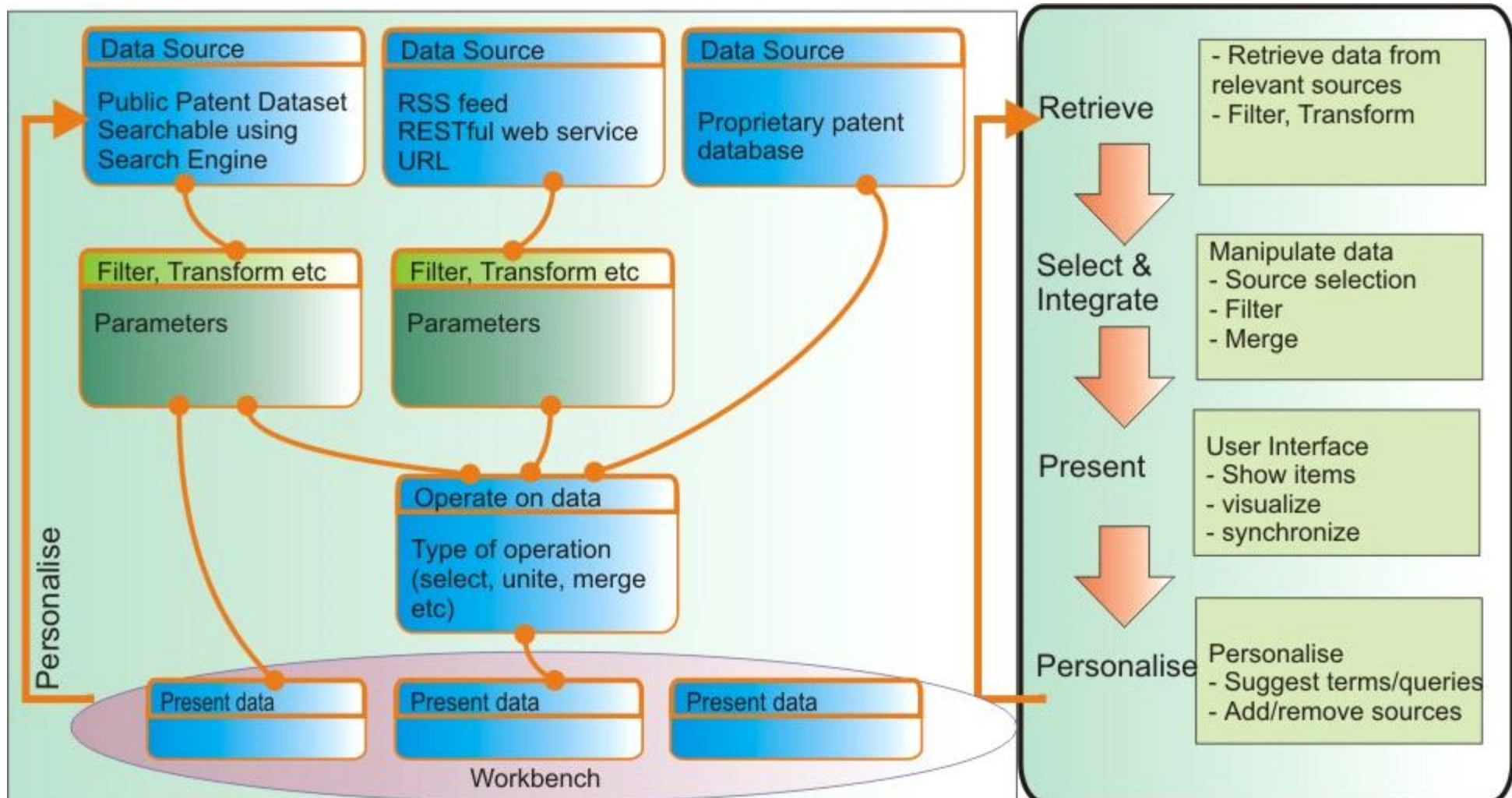
- Integrated search is a methodology utilizing standard search techniques, such as search engines, but integrating multiple sources in the process.
- It may include searching many closely or loosely related databases. However, how closely or loosely related they are depends upon the keywords used.
- An integrated searching capability is also utilized in desktop searching, where it has the ability to simultaneously search hard drives and removable storage on the user's computer.

Definition of Integrated Search Systems

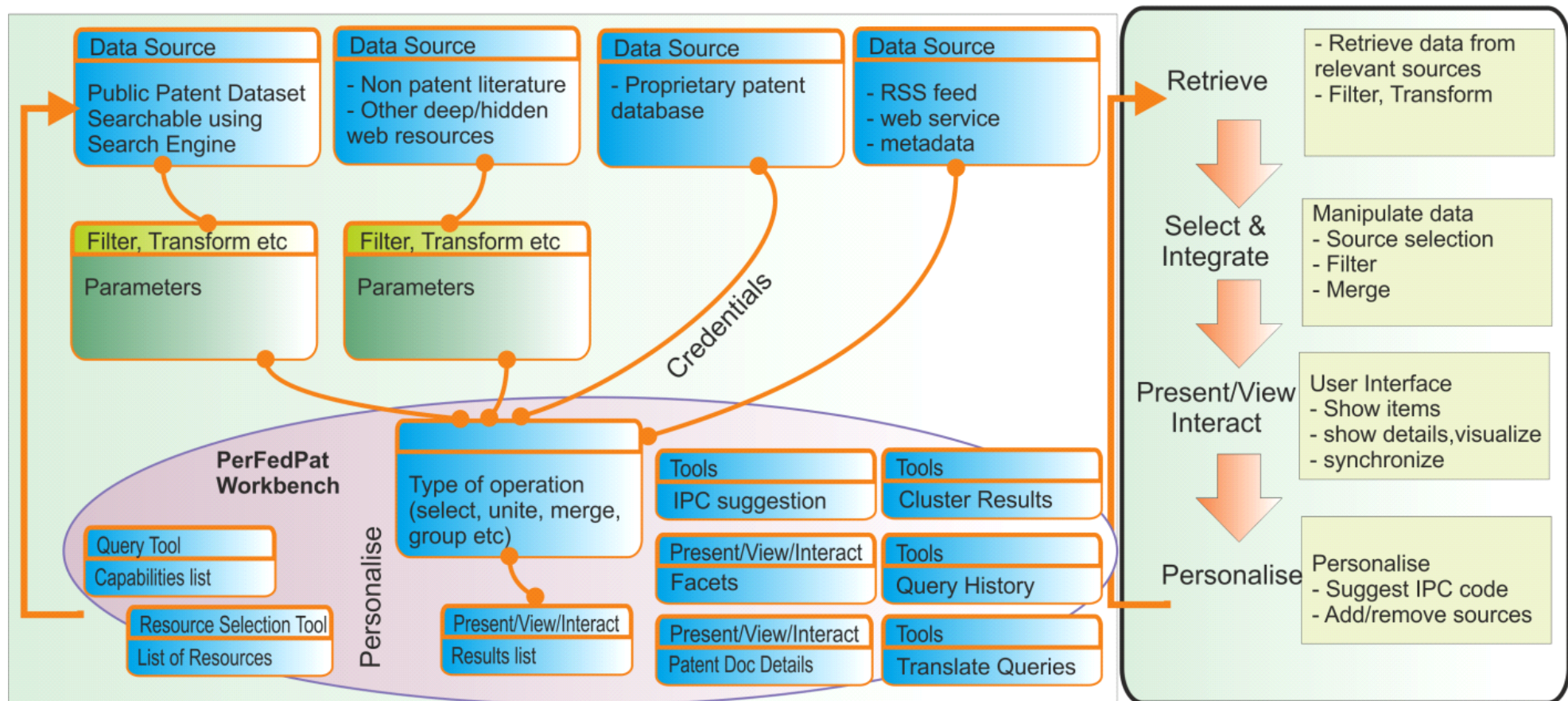
In our definition of *integrated (professional) search systems*,

- the term integrated is used beyond the way that it is used in Federated (or aggregated/integrated) search.
- It is used to define search systems integrating **multiple search tools**
- The tools can co-exist in user's desktop (workbench) and can be used (in parallel or in a pipeline) from the professional searcher during a potentially lengthy search session.

Integrated Search System Architecture



An example: PerFedPat



Interaction Schema

For each *action* in user's Workbench in PerFedPat (e.g. submission of a query)

Repeat

- Retrieve data from N data source(s)

- Transform data appropriately (e.g. translate), select, filter

- Merge data if required

- Present final results, group, visualise etc.

- Notify other search tools and adapt if possible and necessary

Until goal is achieved or search is terminated or saved (user decision)

An example: PerFedPat

The screenshot displays the PerFedPat web application interface, which is a patent search tool. The interface is divided into several panels:

- Advanced Query Panel (A):** Located on the left, it contains search criteria fields for Full Text/Abstract, Title, Publication number, Application number, Priority number, Year, Applicant(s), Inventor(s), European Classification (ECLA), International Patent Classification (IPC), and U.S. Classification. The search term "car AND brake" is entered in the Full Text/Abstract field.
- Library Choice Panel (B):** Located below the Advanced Query Panel, it allows users to select search engines: CIP, Espacenet, G. Patents, and PatentScope.
- Results Panel (C):** Located on the right, it displays a list of search results. The first result is "HAND-OPERATED CAR-BRAKE" by LORD MANUFACTURING CO, with IPCs [B61H13/02] and US-1166214-A, 1915 (Espacenet).
- Cluster Explorer Panel (E):** Located below the Results Panel, it shows a hierarchical view of the search results, including categories like "brake(47)", "freinage(23)", "systeme(20)", "frein(19)", "car(35)", and "control(23)".
- Entities Explorer Panel:** Located to the right of the Cluster Explorer, it shows a list of entities, including "Inventor (125 entities)" and "MCLAUGHLIN BRYAN M US (7)".
- Details Panel (D):** Located at the bottom, it provides detailed information about the selected patent, including the URL "http://ops.epo.org/3.0/rest-services/published-data/search/biblio?q=US1166214" and the title "HAND-OPERATED CAR-BRAKE".

The interface is designed to facilitate patent searches and analysis, providing a comprehensive view of search results and related information.

Motivation for Integrated Search Systems

- There is no IR/NLP technology which can effectively respond to all information needs in all different contexts
- To put it different, despite the tremendous improvement, the search problem is far from solved
- Professional search is much more demanding, many different IR/NLP tools are needed
- Meet and Complement the
 - Open Data,
 - Big Data,
 - Linked Open Data era

Motivation for Integrated Search Systems

- The IR and NLP research communities have achieved tremendous progress in developing new algorithms and tools in various areas of information processing and retrieval,
 - however there was little attention paid on how these results can come together to design next generation search systems.
- This view is supported by the fact that using and managing information workflows between autonomous (and possibly distributed) IR or NLP tools/services is the main design method used by different groups working in managing languages resources or professional search systems.

Develop an Ecosystem for IR/NLP tools to flourish

- Provide a framework to develop an IR/NLP search tools ecosystem where different tools can be straightforward integrated
- Attractive business model for research groups building different types of IR/NLP technologies and tools or for SMEs developing search solutions.

Outline

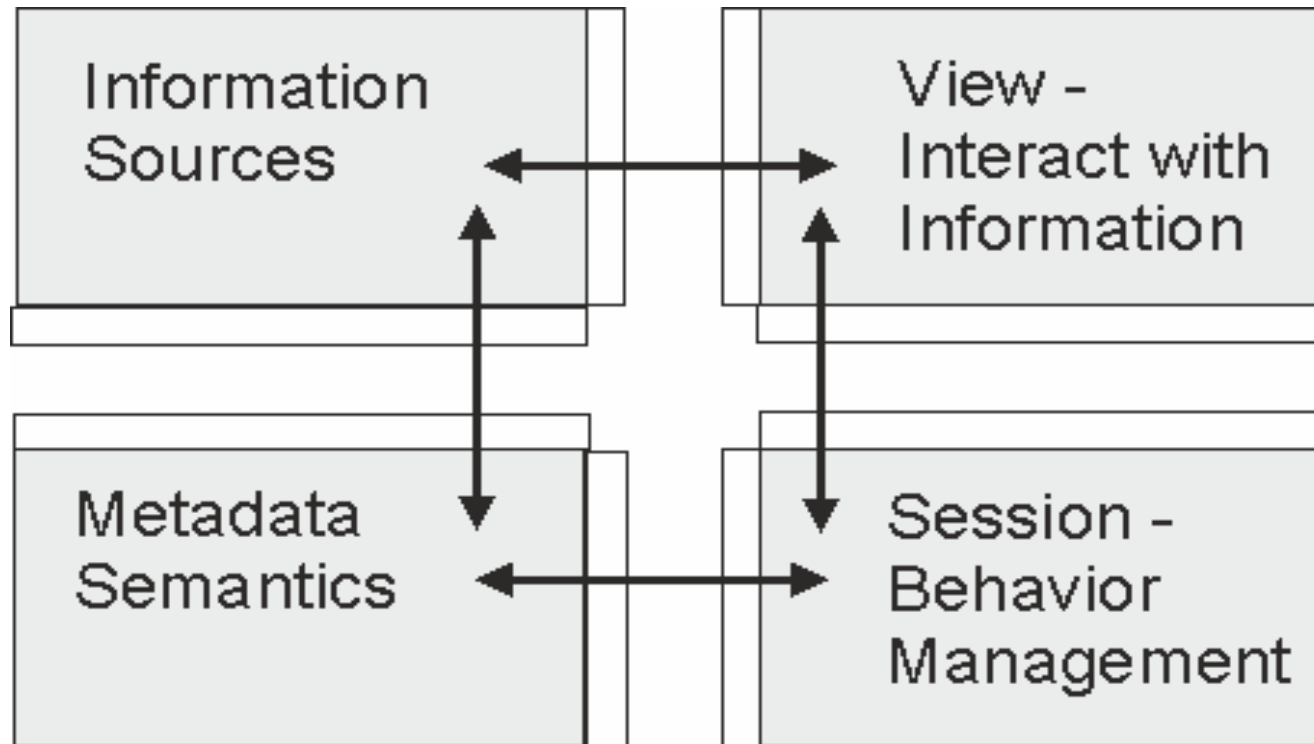


- MUMIA
- Introduction to Professional Search and Some Terminology
- Integrated Search Systems
- **A General Framework for Integrated Professional Search Systems**
- Case Study – Putting things to work
- Open Problems

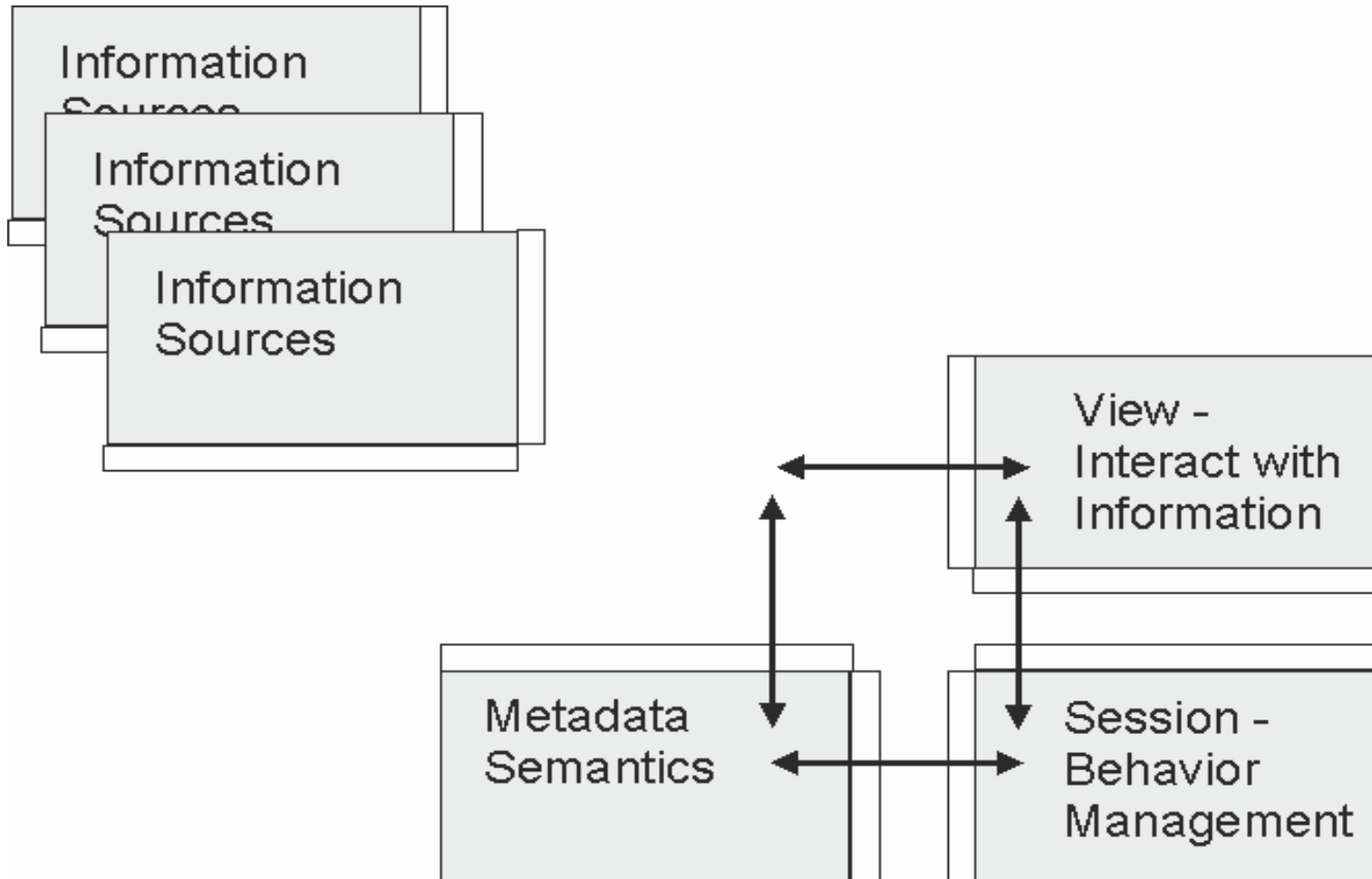
A Framework for Integrated Search Systems

- In this ecosystem we need a method to:
 - better classify and characterize what Integrated Professional Search (IPS) systems are, but
 - better understand the design space of IPS systems
 - describe and compare professional search systems in a more systematic and independent way and,
 - provide an architecture for developing interoperable search systems based on a set of cooperating IR/NLP tools

Taxonomy of an integrated search system

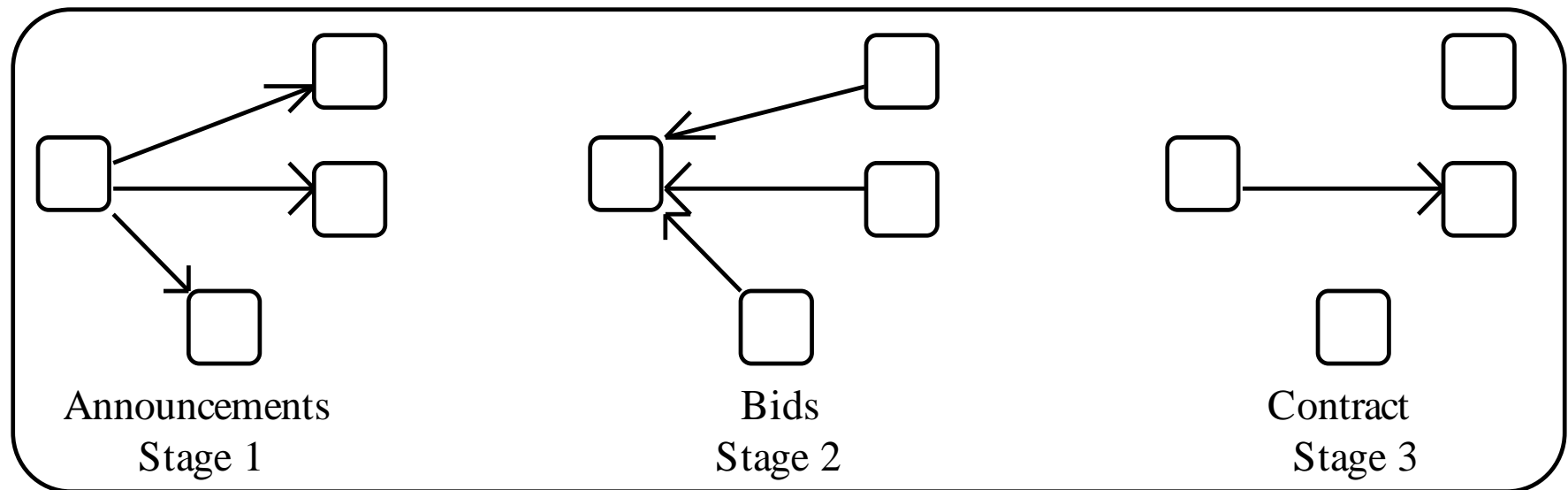


Possibilities/Expresiveness of the taxonomy



Protocols

Communication and Coordination Protocols are required



Outline



- MUMIA
- Introduction to Professional Search and Some Terminology
- Integrated Search Systems
- A General Framework for Integrated Professional Search Systems
- **Case Study – Putting Things to Work**
- Open Problems

Putting the Framework to work

- To evaluate the applicability of the Electra framework we used it during a research networking meeting, where 38 IR/NLP scientists and professionals organised groups participated, in an experiment based on the living lab concept.
- The main purpose was to evaluate the expressiveness of the Electra framework within the context of four different groups:

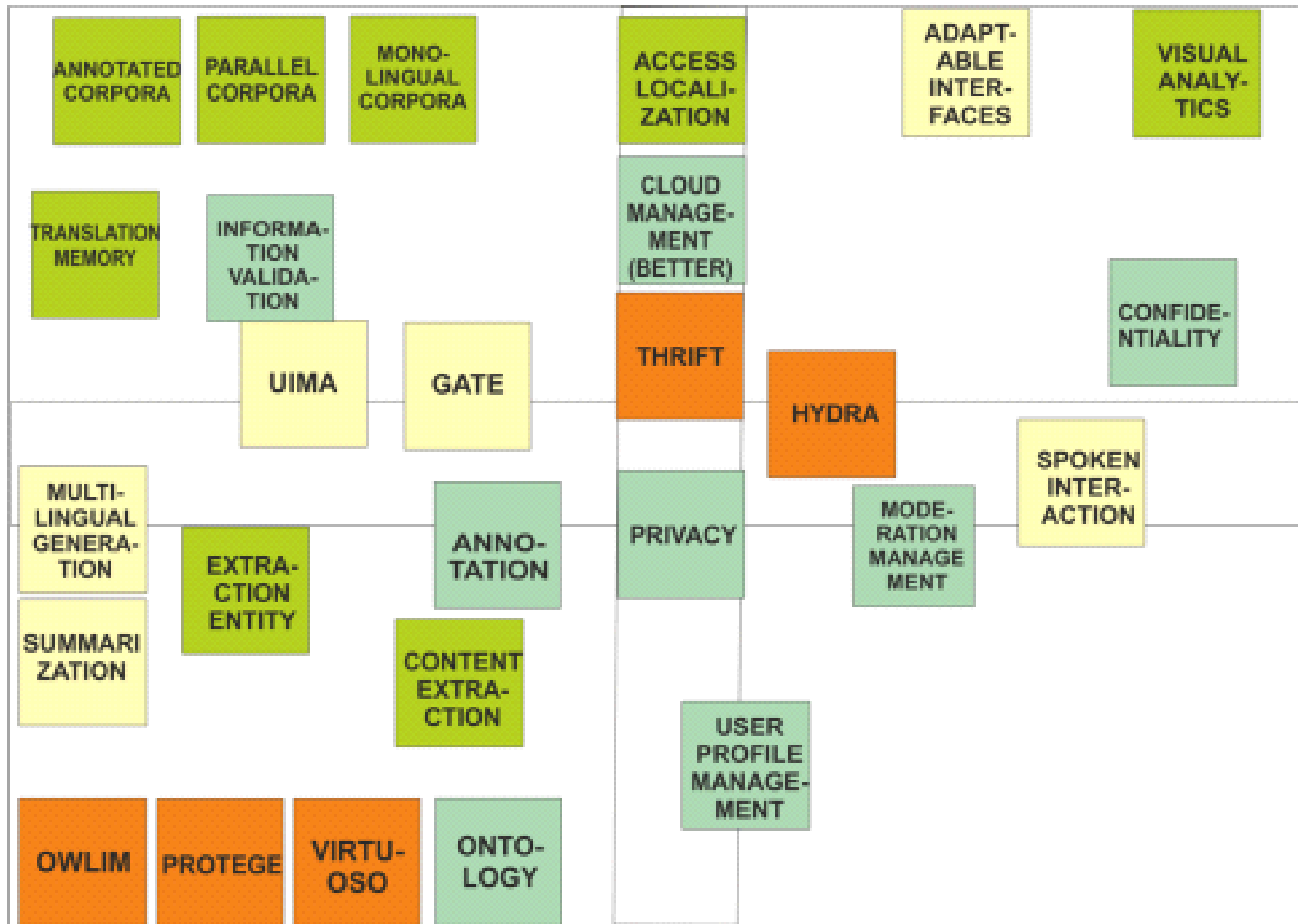
Four different entities were used

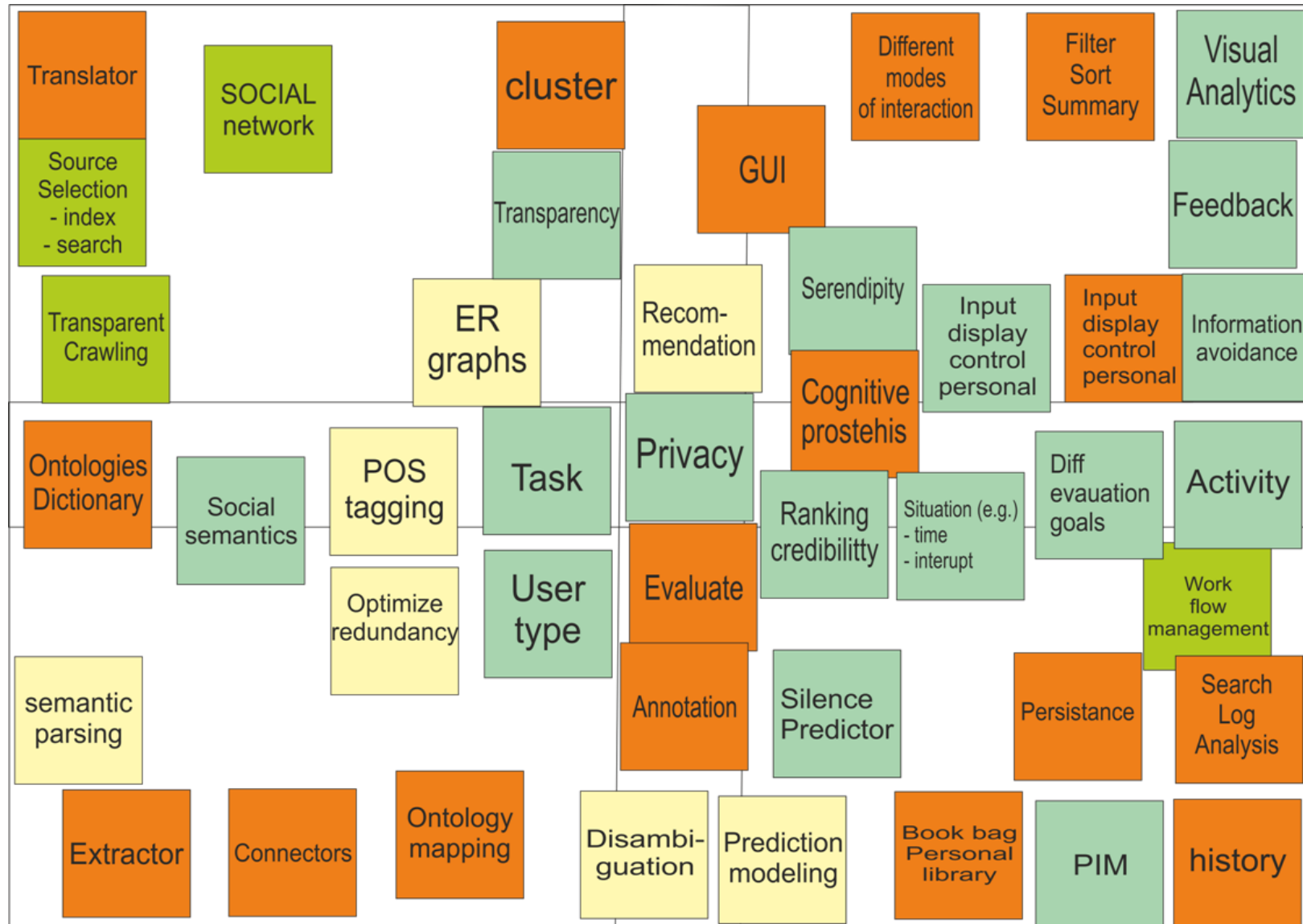
 yellow: IR/NLP technologies

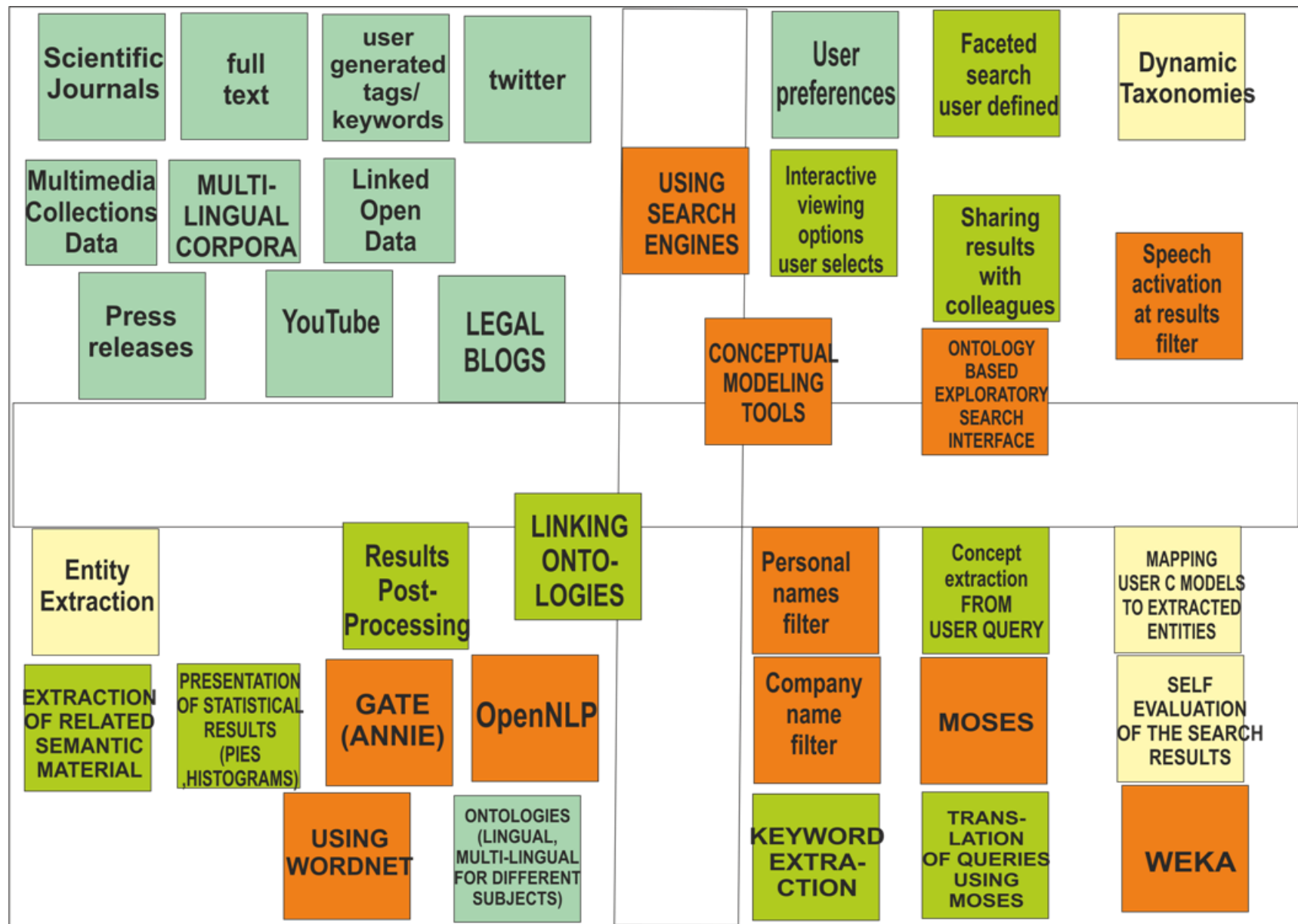
 cyan: concepts

 green: core services

 orange: tools







Which search tools and how should be integrated?

- It is a mistake if we think the search tools which should be integrated into patent search systems depend only on existing IR or text processing technologies,
- Probably it has more to do with the goal of a patent search and the behavior of the searcher.
- Furthermore, it is also very important to deeply understand a search process and how a specific tool can attain a specific objective of this process and therefore increase its efficiency.

Understanding Patent Search processes*

1. User receives a patent application document to evaluate
2. User enters the search system
3. User enters a text query, potentially with Boolean operators and specific field filters
4. System presents a result list, sorted by relevance, with snippets, metadata information, and links to full documents
5. User inspects and assesses all the documents
6. User clicks on one element of the list for further inspection
7. System presents the full document, with any metadata, attached images and text
8. User inspects and assesses. Finds the document potentially relevant and saves it to a bucket
9. User clicks on the 'Back' button to return to the list of results
10. System presents the list, with the already viewed documents visibly identifiable
11. Jump to Step 6, unless new query query is required or User satisfied
12. Based on a potential new understanding, User inputs a new query
13. Jump to Step 3, unless User satisfied
14. User saves the list and creates a search report
15. Use case ends

* Taken from Mihai Lupu and Allan Hanbury, Review Patent Retrieval

MULTILAYER COLLECTION SELECTION AND SEARCH OF TOPICALLY ORGANIZED PATENTS



Topically Organised Patents

Table 1. An Example of a Section From the IPC Clasification

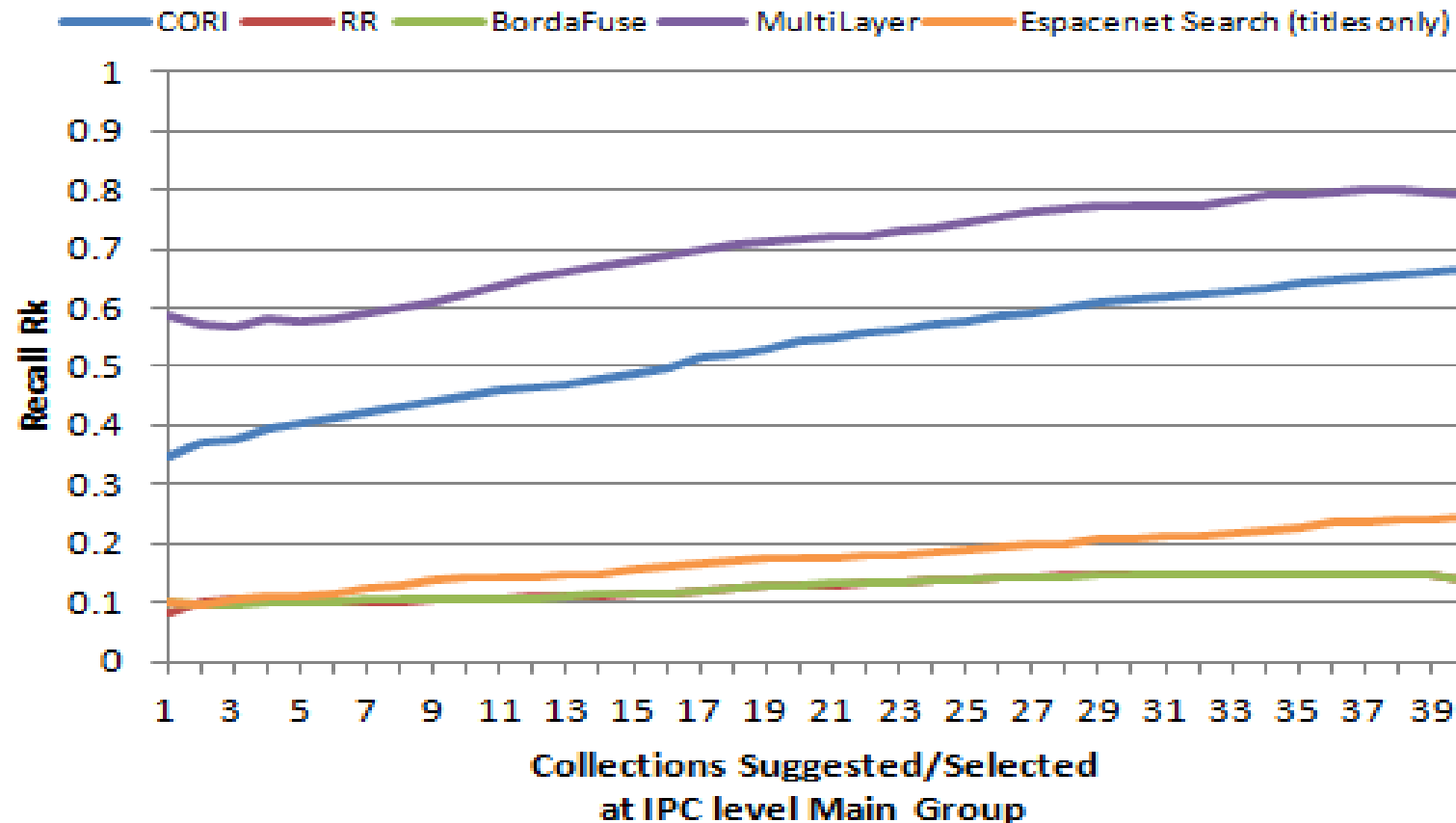
Section	Mechanical engineering...	F
Class	Machines or engines in general	F01
Subclass	Machines or engines with two or more pistons	F01B7
Main group	reciprocating within same cylinder or ...	F01B7/00
Subgroup	.with oppositely reciprocating pistons	F01B7/02
Subgroup	..acting on same main shaft	F01B7/04

Topically Organised Patents

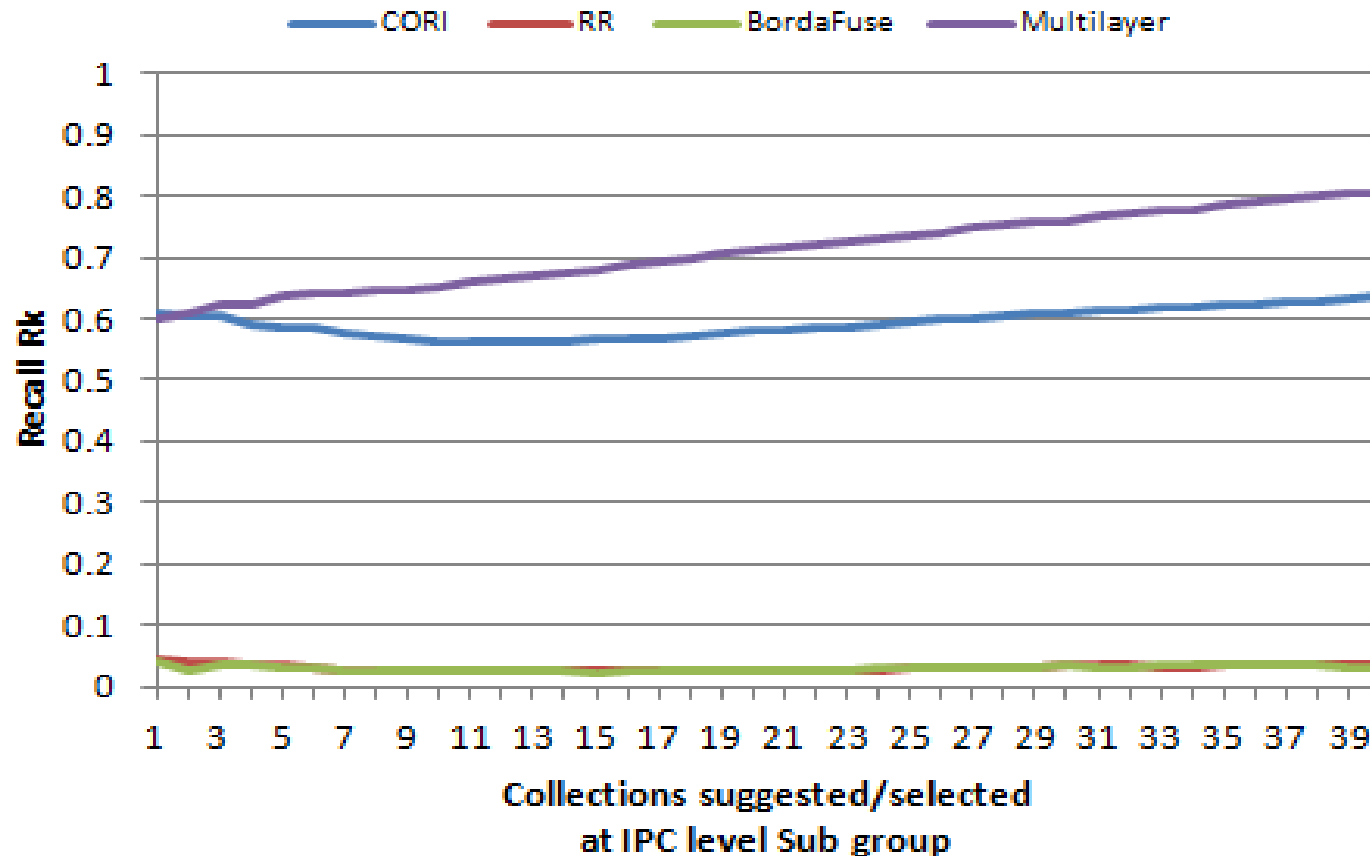
Table 2. Statistics of the CLEF-IP 2012 divisions using different levels of IPC

Split	# patents	Collections Number	Docs per collection			
			<u>Avg</u>	Min	Max	Median
split_3	3622570	632	5732	1	165434	1930
split_4	5363045	7530	712	1	83646	144
split_5	10393924	63806	163	1	39108	36

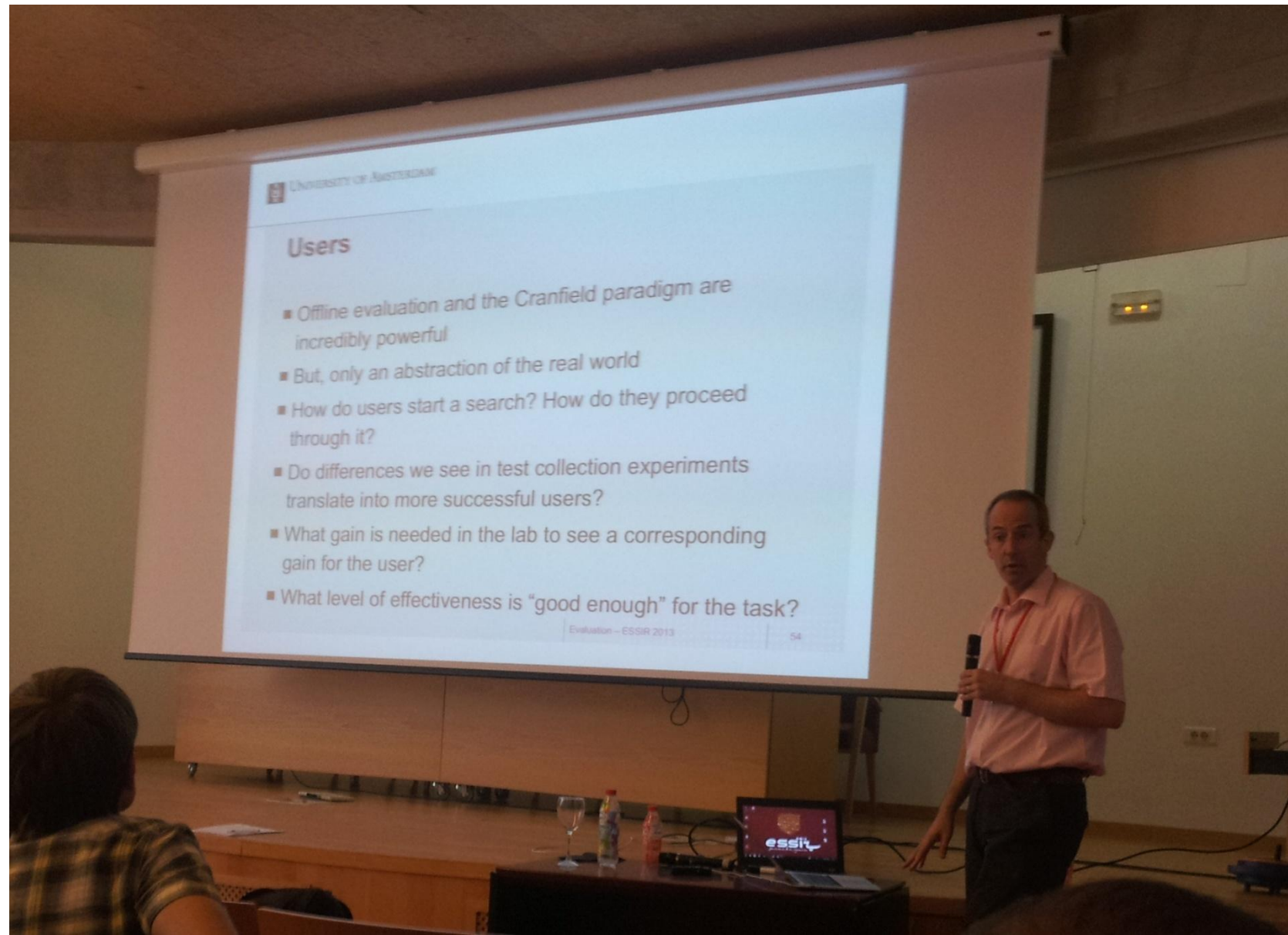
Source Selection Results (level 4)



Source Selection Results (level 5)



“Do differences we see in test collections translate into more successful users?”, from Maarten’s talk



How do we integrate the IPC suggestion tool?

The screenshot displays the ezDL web interface, which is used for patent searches. The interface is divided into several sections:

- Advanced Query:** This section contains search criteria for a patent query. The search terms are "car AND brake". Other criteria include Title: "e.g. plastic AND bicycle", Publication number: "e.g. WO2008014520", Application number: "e.g. DE19971031696", Priority number: "e.g. WO1995US15925", Year: "e.g. >=2005 AND <=2009", Applicant(s): "e.g. 'Institut Pasteur'", Inventor(s): "e.g. Smith", European Classification (ECLA): "e.g. F03G7,10", and International Patent Classification (IPC): "e.g. H03M1,12". The U.S. Classification is "e.g. 280,251". There are "Clear" and "Search" buttons at the bottom of this section.
- Library Choice:** This section allows users to select the patent database to search. The "Patent Resources" dropdown is set to "Clef-IP". Other options include "Espacenet", "G. Patents", and "PatentScope".
- IPC Suggestions:** This section displays a list of suggested IPC classes based on the search criteria. The suggestions are: 1. Suspension system for a car mounted brake assembly (WESTINGHOUSE AIR BRAKE CO), 2. Integrated train electrical and pneumatic brakes (NEW YORK AIR BRAKE CORP), 3. RAIL CAR LOAD SENSOR (TECH SERV & MARKETIN & INC), and 4. Hand brake for a rail car (JACKSON JOHN M; JACKSON ROBERT G).
- Results:** This section shows the search results. The "Results: 200" are displayed. The "Group by" dropdown is set to "Nothing". The "Relevance" dropdown is set to "Relevance". The "Ascending" checkbox is checked.
- Details:** This section provides detailed information about the selected patent. The title is "Suspension system for a car mounted brake assembly". The year is 2002. The IPC classes are B61H1/00, B61H13/00, and B61H13/36. The IPC class H03M1/12 is also shown.

The interface is running on a Windows 7 desktop environment, as indicated by the taskbar and the system clock showing 12:44 μμ Δευτέρα 2/9/2013.

How do we integrate the IPC suggestion tool?

The screenshot displays the ezDL web application interface, which is designed for patent searching and integration with the IPC suggestion tool. The interface is divided into several main sections:

- Advanced Query:** This section contains a search bar with the text "car AND brake". Below the search bar, there are buttons for "Clear" and "Search". To the right of the search bar, there is a button labeled "IPC Suggestions" and a radio button selection for "Level 3", "Level 4", and "Level 5".
- Library Choice:** This section allows users to select the database they want to search. It includes a "Refresh" button, a "Patent Resources" dropdown menu, and a search filter input. The selected database is "Clef-IP", and the search is performed on the "Clef-IP 2011 collection". Other options include "Espacenet", "G. Patents", and "PatentScope".
- Results:** This section displays the search results. It shows a list of 51 results, with the first four results visible. Each result includes a title, a brief description, and the IPC class number. The results are sorted by "Relevance" and "Group by: Nothing".
- Details:** This section provides detailed information about the selected patent, including the title, inventor, and IPC class number.

The interface also features a sidebar on the left with various navigation links and a bottom status bar showing the search progress and time.

Search Results:

- [B60L]PROPULSION OF ELECTRICALLY-PROPELLED VEHICLES (arrangements or mounting of electrical propulsion)**
IPCs: [B60L] 2013 (Clef-IP IPC)
- [B61L]GUIDING RAILWAY TRAFFIC; ENSURING THE SAFETY OF RAILWAY TRAFFIC (brakes or auxiliary equipment B61L)**
IPCs: [B61L] 2013 (Clef-IP IPC)
- [B60T]VEHICLE BRAKE CONTROL SYSTEMS OR PARTS THEREOF; BRAKE CONTROL SYSTEMS OR PARTS THEREOF,**
IPCs: [B60T] 2013 (Clef-IP IPC)
- [B66B]ELEVATORS; ESCALATORS OR MOVING WALKWAYS (life-saving devices used as an alternative to normal eg)**
IPCs: [B66B] 2013 (Clef-IP IPC)

Search Results (Continued):

- Suspension system for a car mounted brake assembly**
WESTINGHOUSE AIR BRAKE CO
IPCs: [B61H1/00, B61H13/00, B61H13/36] EP-1092607-A3, 2002 (Clef-IP)
- Integrated train electrical and pneumatic brakes**
NEW YORK AIR BRAKE CORP
IPCs: [B60T13/66, B60T17/22, B60T17/18] EP-1084925-A3, 2002 (Clef-IP)
- RAIL CAR LOAD SENSOR**
TECH SERV & MARKETIN & INC
IPCs: [B60T13/66, B60T8/18, B60T8/22, B60T] WO-1997018979-A1, 1997 (Clef-IP)
- Hand brake for a rail car**
JACKSON JOHN M; JACKSON ROBERT G
IPCs: [B60T7/10, B61H13/02, B60T7/02, B61H13/00] EP-1177964-A3, 2002 (Clef-IP)

Details:

ezDL
Welcome to ezDL

Outline



- MUMIA
- Introduction to Professional Search and Some Terminology
- Integrated Search Systems
- A General Framework for Integrated Professional Search Systems
- Case Study – Putting Things to Work
- **Open Problems**

Open Problems



- **Define protocols and standards to facilitate the development of the ecosystem**
- **Harmonise research within the broader context of search systems development**
- **Think beyond the PhD research objective: “I must do an experiment which will show 5% improvement”**

Thank you



*Multilingual and Multifaceted Interactive
Information Access*