Search
Explore
Modify Engine

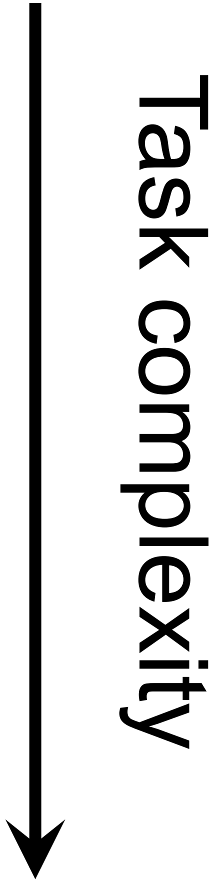
Arjen P. de Vries
arjen@acm.org

Centrum Wiskunde & Informatica
Delft University of Technology
Spinque B.V.

Library of the “Muntmuseum” in Utrecht (Erik van Hannen)

Search Intermediaries

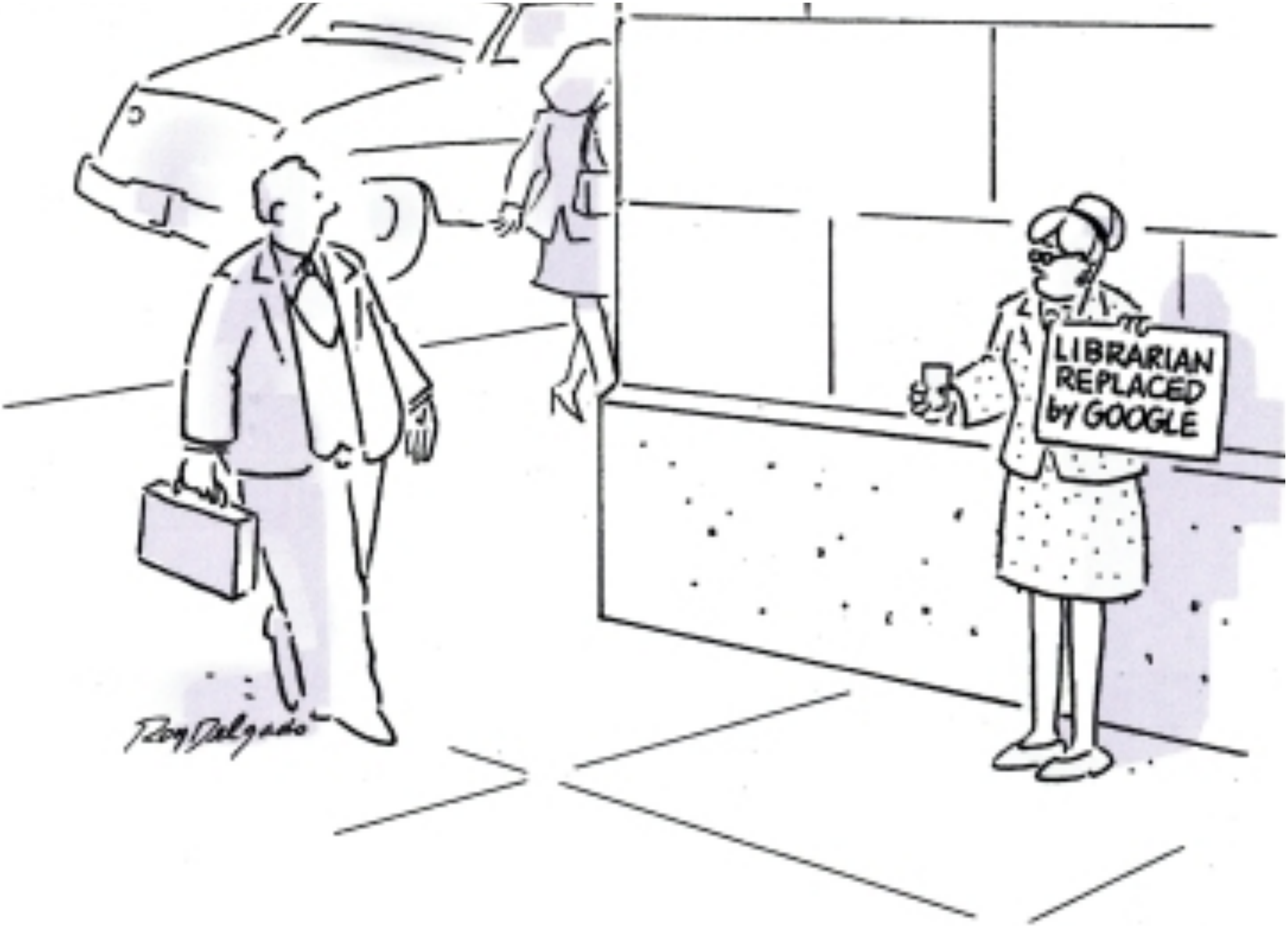
- Travel agency
- Real estate agents
- Recruiters
- Librarians
- Archivists
- Digital forensics detectives
- Patent information specialists

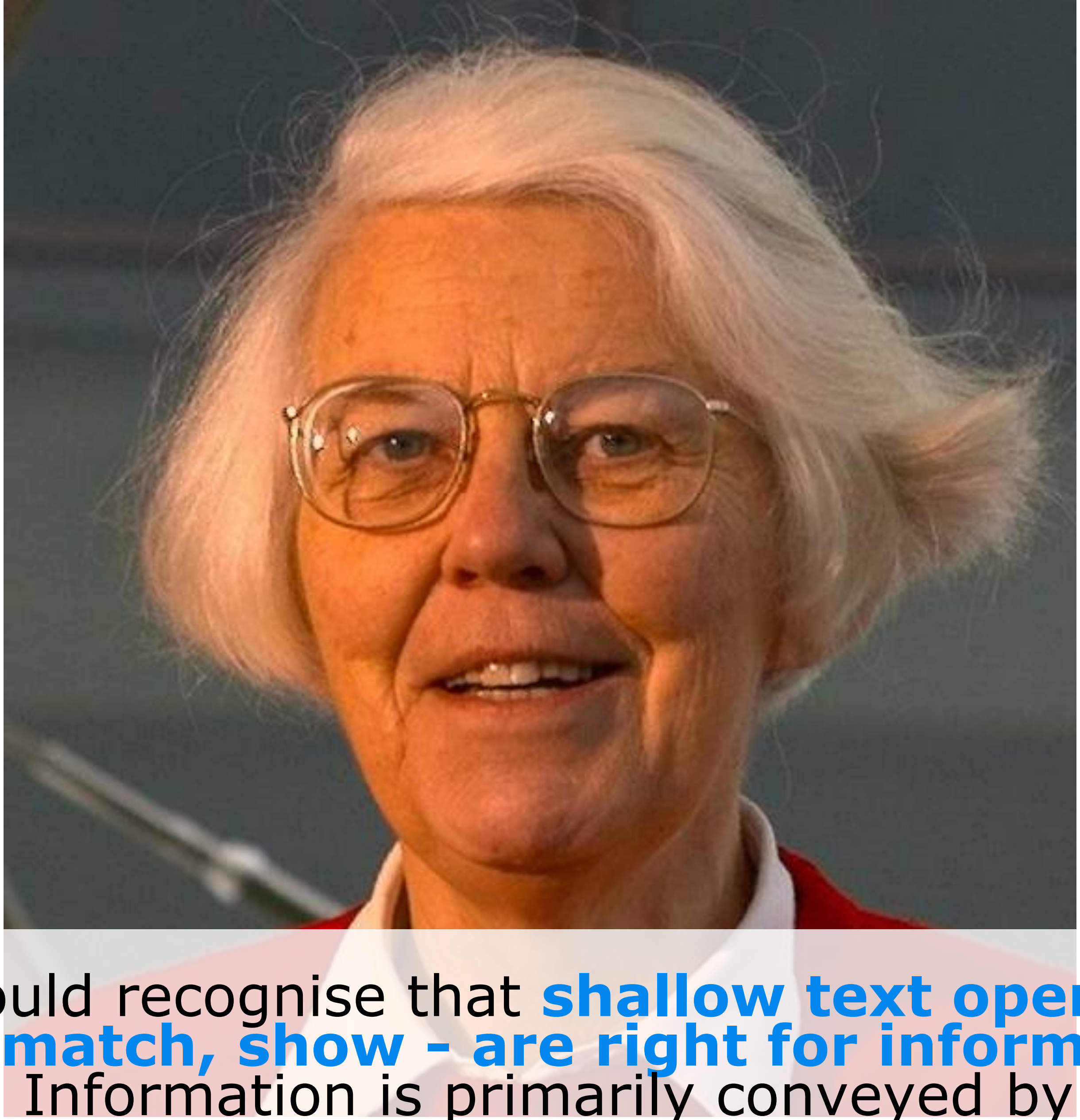


Task complexity

Trend

- Do-It-Yourself (DIY) information seeking
  - *Convenient* access to online search engines
  - *Perceived* time efficiency


LIBRARIAN
REPLACED
by GOOGLE
Roy Delgado



"We should recognise that **shallow text operations - select, match, show - are right for information access.** Information is primarily conveyed by natural language and this has to be shown to the user for them to assess."

Karen Sparck Jones. *What's new about the Semantic Web? Some questions.* In SIGIR Forum, Volume 38 Issue 2, December 2004

Trend

- Do-It-Yourself (DIY) information seeking
  - *Convenient* access to online search engines
  - *Perceived* time efficiency
- Let’s face it:
  - Google/Bing/Y! is often best
  - *Even Google Enterprise Search (“the Google Box”) is far worse than Google Web Search!*

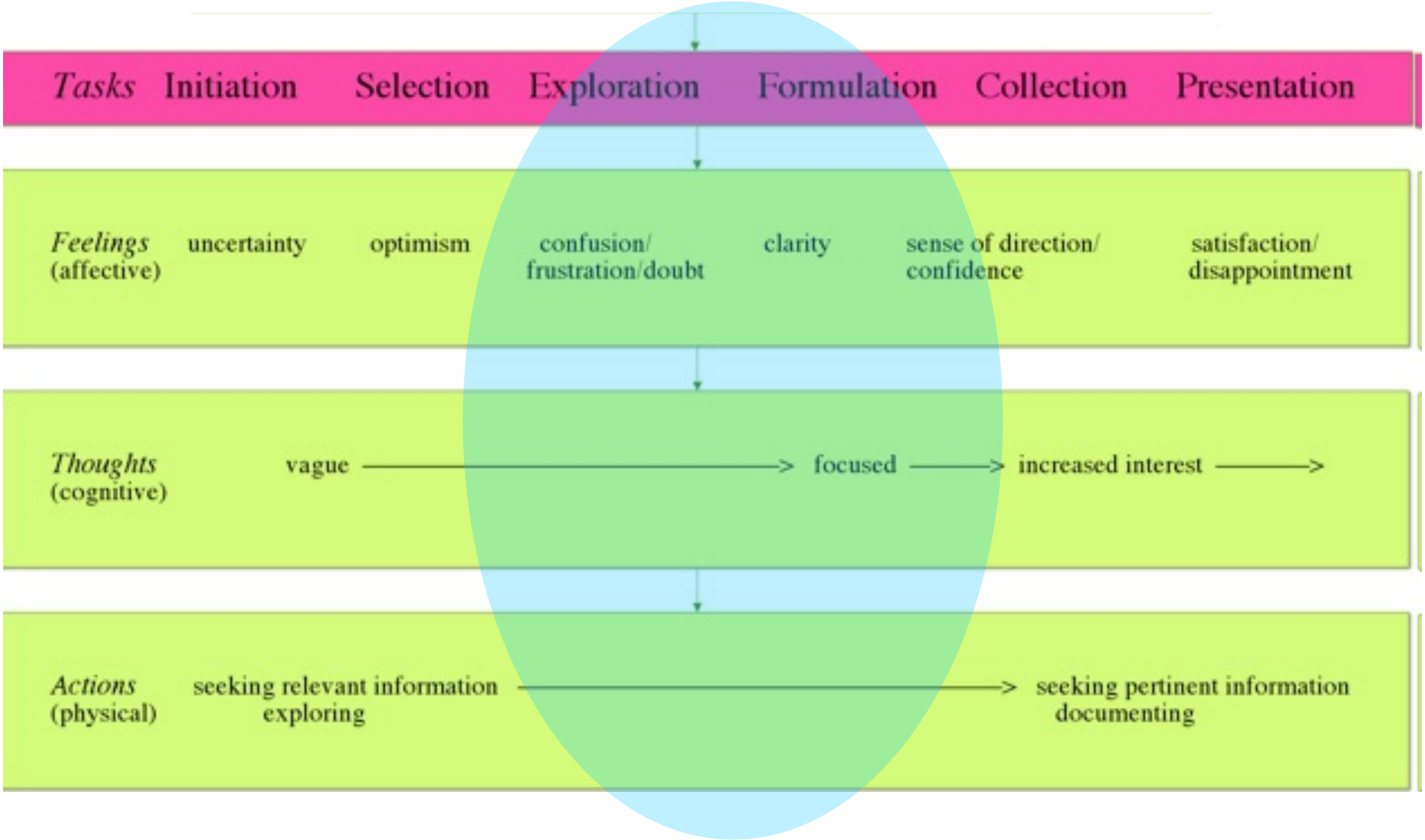
Kuhlthau six stages

- Initiation: user "becomes aware of a lack of knowledge or understanding"
- Selection: user needs to "identify and select the general topic to be investigated"
- Exploration: user needs to "investigate information on the general topic in order to extend personal understanding"
- Formulation: user forms "a focus from the information encountered"
- Collection: user needs "to gather information related to the focused topic"
- Presentation: user completes the search and presents findings

Exploration, Formulation

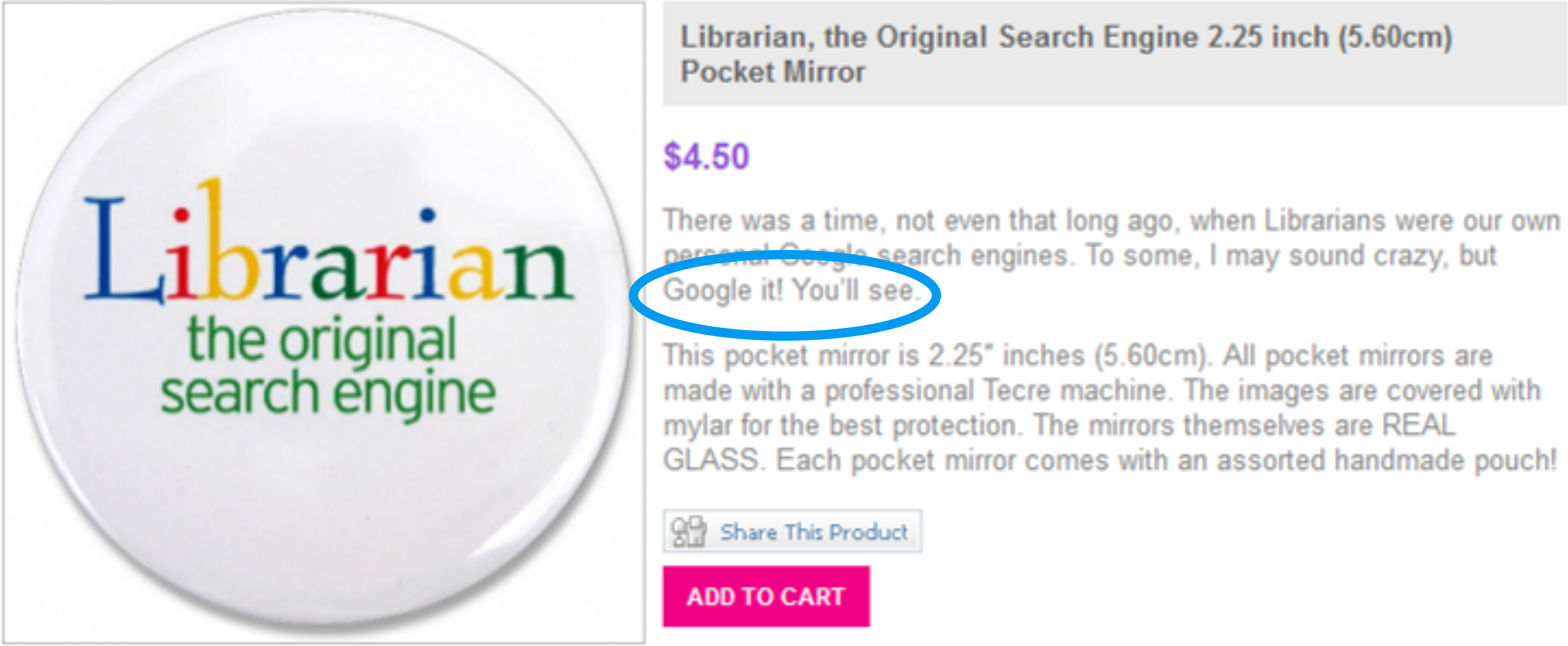
- *I want to buy a house in Amsterdam and I want it with 'sfeer' but still in good shape*
- *I can afford about €350K. I need 3 bedrooms, the size should be about 80m^2. It should have a balcony or a backyard*
- *The closer to the station and an AH, the better. BUT… I do not want to live in Amsterdam-Noord, unless there is a quick bus connection to the ferry*
- *I may be willing to drop some of these constraints, but I'm not sure which*

Seeking Search Intermediary?!



Tasks	Initiation	Selection	Exploration	Formulation	Collection	Presentation
Feelings (affective)	uncertainty	optimism	confusion/ frustration/doubt	clarity	sense of direction/ confidence	satisfaction/ disappointment
Thoughts (cognitive)	vague →			focused →	increased interest →	
Actions (physical)	seeking relevant information exploring →					seeking pertinent information documenting

Formative Stages of the Information Seeking Process



Librarian, the Original Search Engine 2.25 inch (5.60cm) Pocket Mirror

$4.50

There was a time, not even that long ago, when Librarians were our own personal Google search engines. To some, I may sound crazy, but Google it! You'll see.

This pocket mirror is 2.25" inches (5.60cm). All pocket mirrors are made with a professional Tecre machine. The images are covered with mylar for the best protection. The mirrors themselves are REAL GLASS. Each pocket mirror comes with an assorted handmade pouch!

Disclosure: I have been a librarian!

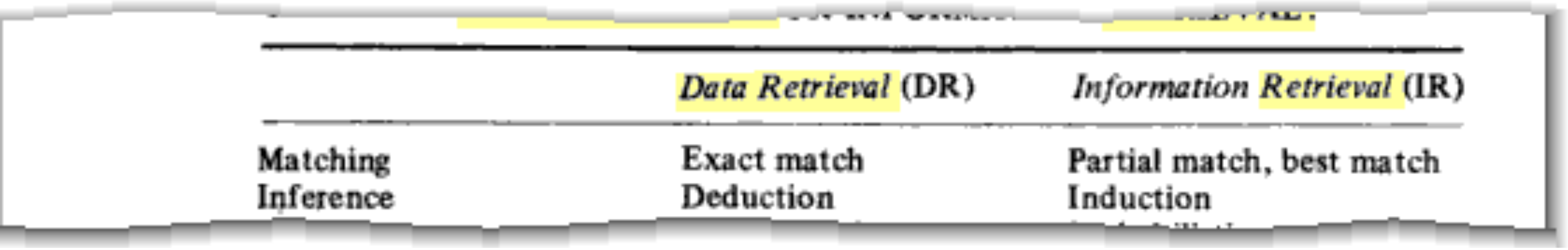
Library of the University of Utrecht (Erik van Hannen)

Trend

- Do-It-Yourself (DIY) information seeking
  - *Convenient* access to online search engines
  - *Perceived* time efficiency
- Let's face it:
  - Google/Bing/Y! is often best
  - *Even Google Enterprise Search ("the Google Box") is far worse than Google Web Search!*
- Lack of tools for the search intermediary to do better than Google?!

Search = IR + DB

- Search tasks in the formative stages of ISP are likely to benefit from
  - a mix of exact (DB) and ranked (IR) searches
  - on structured (DB) and unstructured (IR) data
- Current technical solutions support either/or
- Combining results requires significant effort
  - copy & paste result sets between interfaces, “human (probabilistic) joins”



	Data Retrieval (DR)	*Information Retrieval* (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction

Van Rijsbergen, 1979

Search = IR on-top-of DB ?

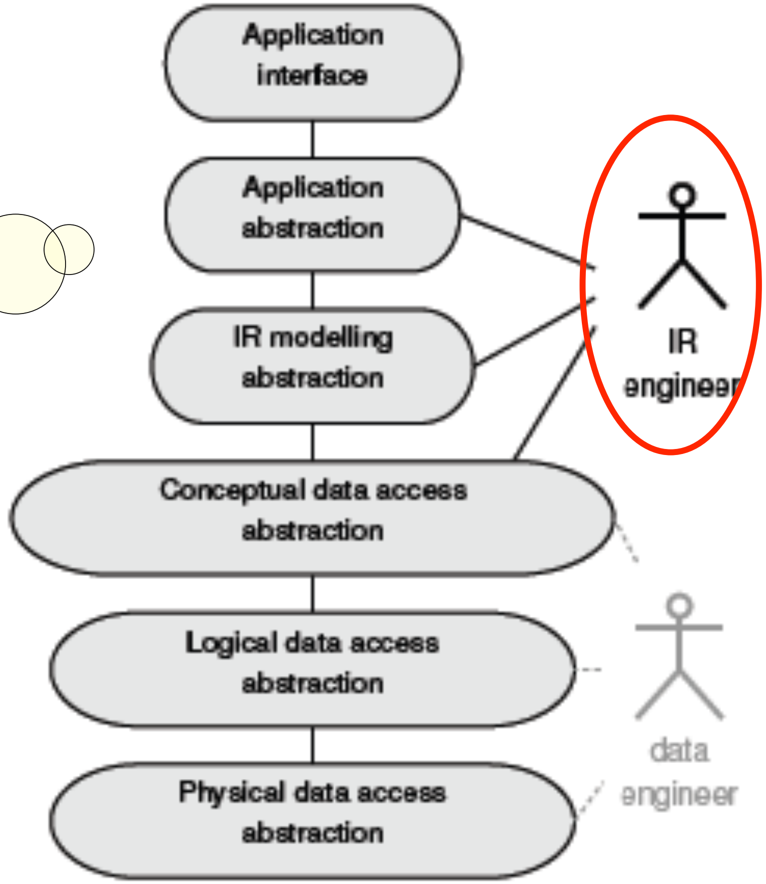
- IR on-top-of DB: let exact and ranked operations both be processed by the same engine, so they can be mixed freely
- IR responsible for ranking models, using DB as a data-access layer; no physical details necessary
- DB responsible for reliable, dynamically optimised, data access; no logical details necessary

IR on-top-of DB???!

- Traditional, general-purpose DB technology cannot compete with custom IR search tools
    - Working assumption: using column stores should solve the efficiency problem

Parameterised Search System (PSS)

Cannot we 'remove' this IR engineer from the loop, like DBMS software removes the data engineer from the loop?


Application interface
Application abstraction
IR modelling abstraction
Conceptual data access abstraction
Logical data access abstraction
Physical data access abstraction
IR engineer
data engineer

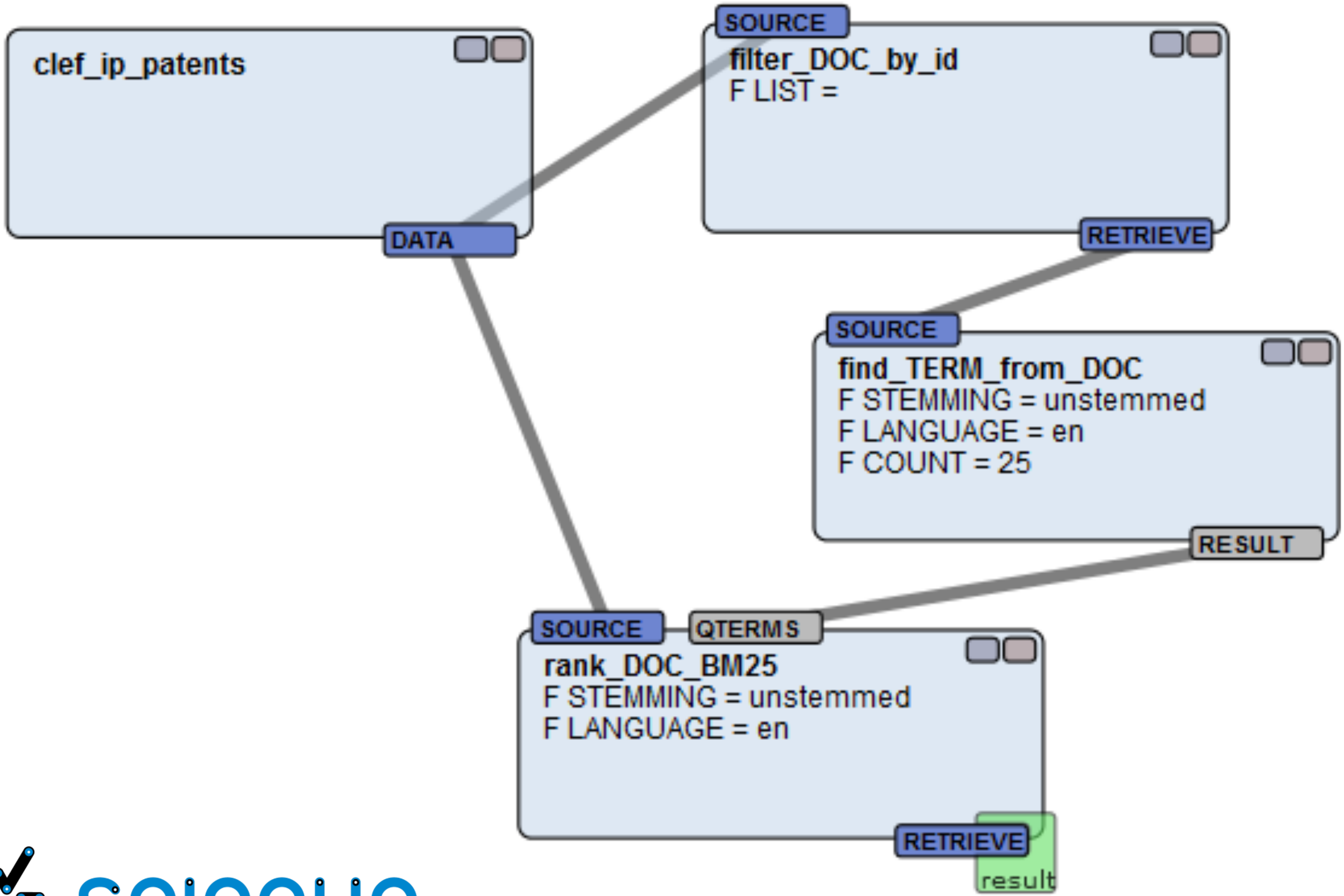
(b) Parametrised, IR and data engineering are two separate roles (possibly automated data engineering in grey)

Cornacchia, De Vries, ECIR 2007
A Parametrised Search System

Search
Extract
TFIDF
results
Select
additional
terms
New
search
STAPLES

Search by Strategy

- Visually construct search strategies by connecting building blocks

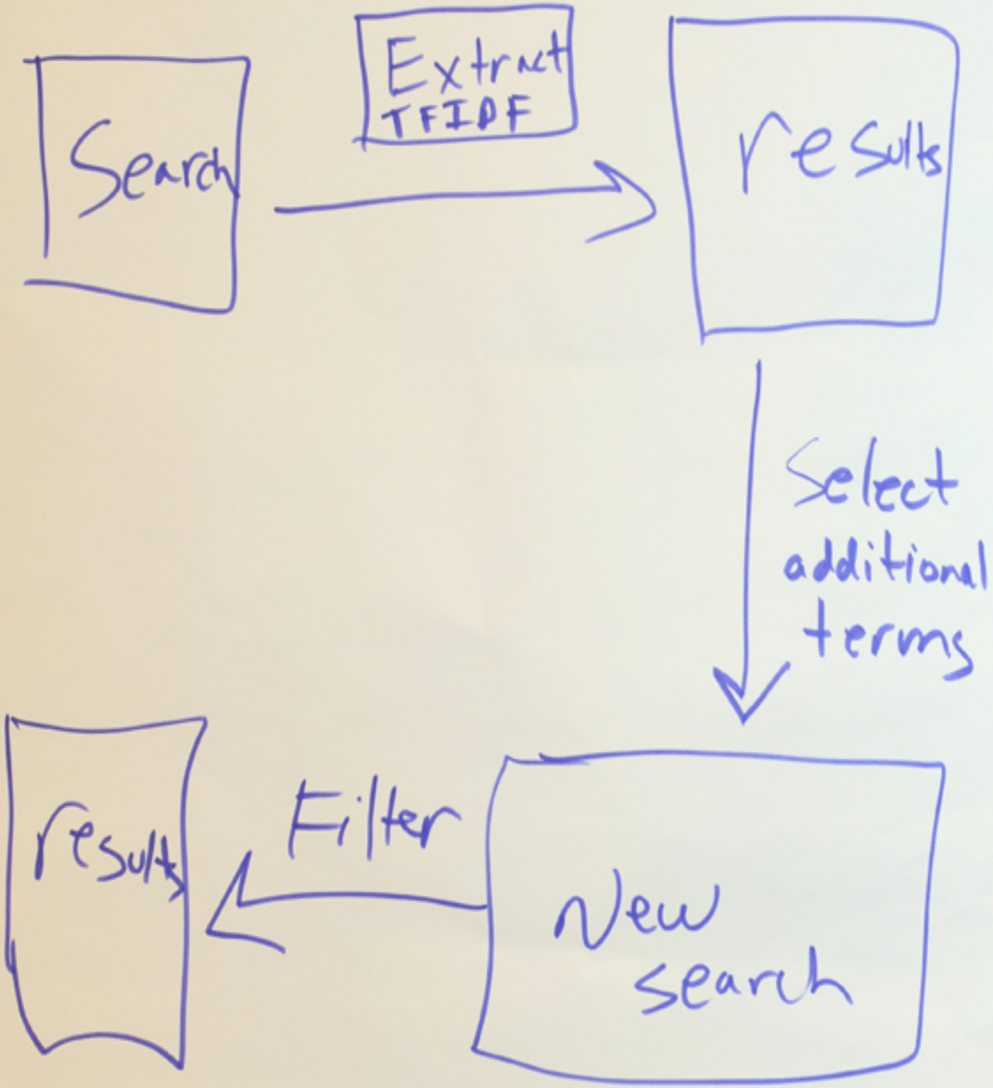


clef_ip_patents

filter_DOC_by_id
F LIST =

find_TERM_from_DOC
F STEMMING = unstemmed
F LANGUAGE = en
F COUNT = 25

rank_DOC_BM25
F STEMMING = unstemmed
F LANGUAGE = en



Need building blocks which correspond to frequentl used task (example a synonym builder which potentially communitydrive

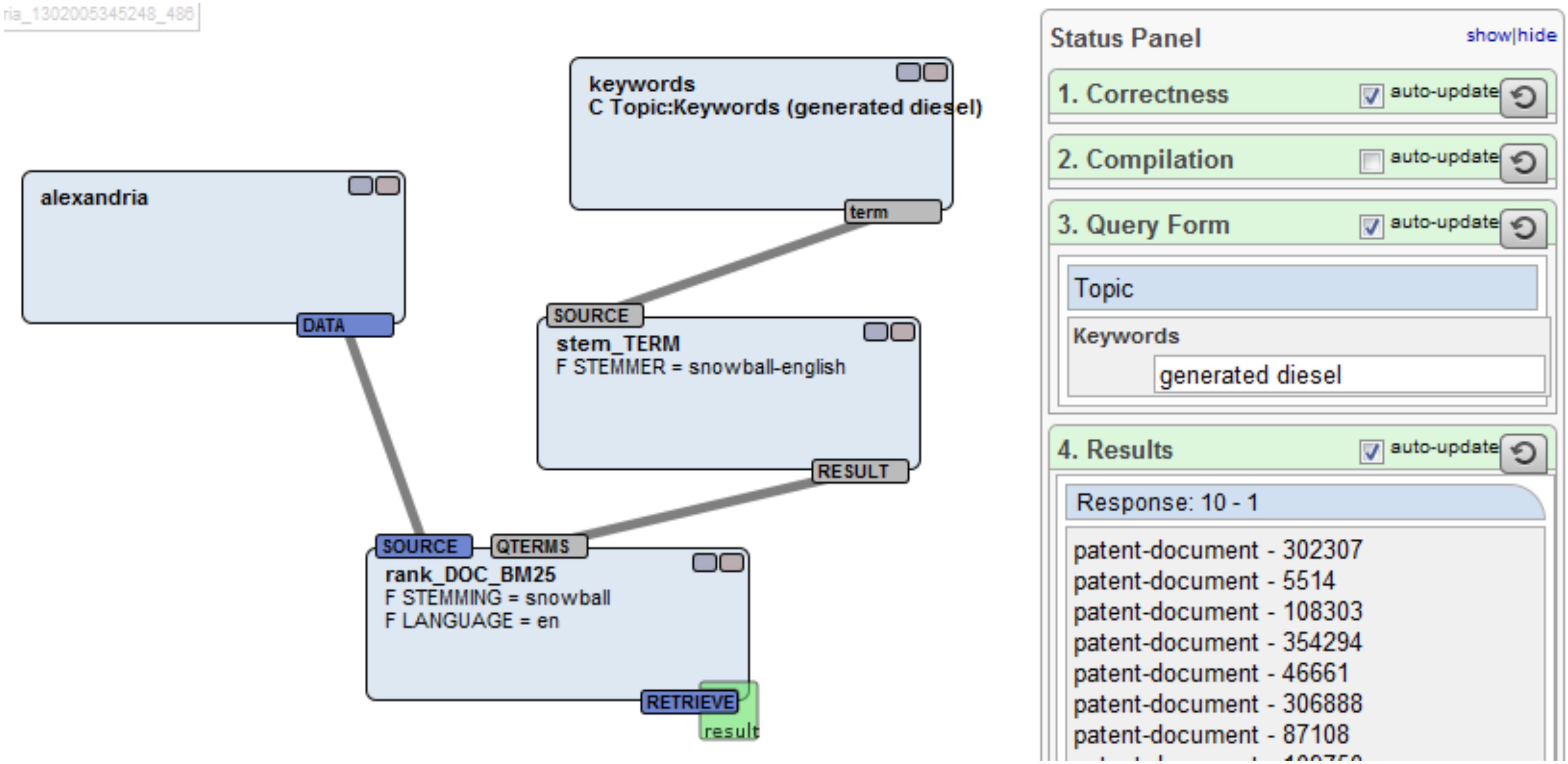
Open Source ontologi maybe available

20 sheets Blatt hojas STAPLES
20 sheets Blatt hojas STAPLES

Search by Strategy

- Visually construct search strategies by connecting building blocks
- Each block describes either data or actions upon that data

Strategy Builder



alexandria

keywords
C Topic:Keywords (generated diesel)

stem_TERM
F STEMMER = snowball-english

rank_DOC_BM25
F STEMMING = snowball
F LANGUAGE = en

Status Panel

1. Correctness ☑ auto-update

2. Compilation □ auto-update

3. Query Form ☑ auto-update

Topic

Keywords: generated diesel

4. Results ☑ auto-update

Response: 10 - 1

patent-document - 302307
patent-document - 5514
patent-document - 108303
patent-document - 354294
patent-document - 46661
patent-document - 306888
patent-document - 87108

Search by Strategy

- Data sources are internally represented as quadruples, triples extended with an additional probability value
- Actions are scripts expressed in (a variant of) Fuhr and Roelleke’s PRA (TOIS 1997)
  - Boolean search: limit probabilities to 0 and 1!
- A search strategy may include multiple data sources

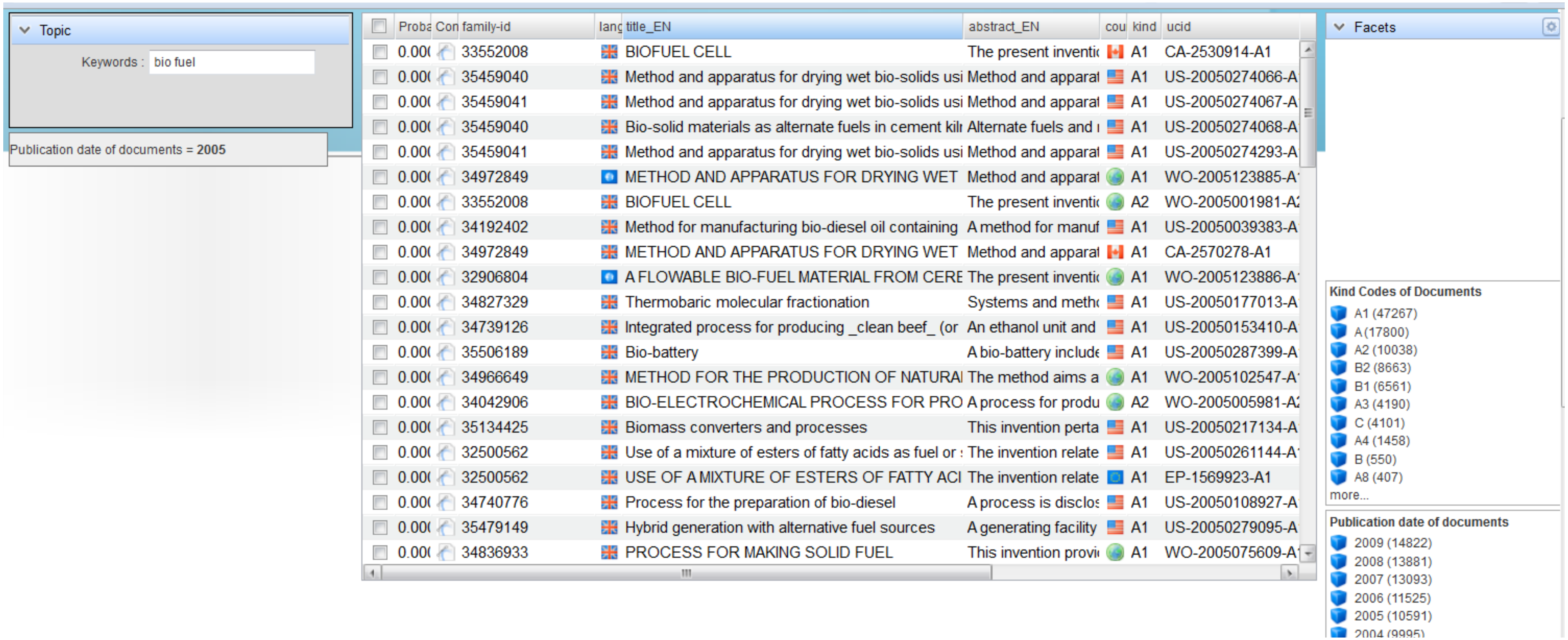
Implementation

- PRA translates into SQL (!)
- Current system setup using CWI’s MonetDB column-store
- Strategies are dynamically transformed into a REST API *and* a GWT UI

monetdb

Beeld en Geluid

Generate Search Engine!



Topic

Keywords : bio fuel

Publication date of documents = **2005**

□	Proba	Con	family-id	lang	title_EN	abstract_EN	cou	kind	ucid
□	0.00(		33552008	🇬🇧	BIOFUEL CELL	The present inventi	🇨🇦	A1	CA-2530914-A1
□	0.00(		35459040	🇬🇧	Method and apparatus for drying wet bio-solids usi	Method and apparat	🇺🇸	A1	US-20050274066-A
□	0.00(		35459041	🇬🇧	Method and apparatus for drying wet bio-solids usi	Method and apparat	🇺🇸	A1	US-20050274067-A
□	0.00(		35459040	🇬🇧	Bio-solid materials as alternate fuels in cement kil	Alternate fuels and	🇺🇸	A1	US-20050274068-A
□	0.00(		35459041	🇬🇧	Method and apparatus for drying wet bio-solids usi	Method and apparat	🇺🇸	A1	US-20050274293-A
□	0.00(		34972849	🌐	METHOD AND APPARATUS FOR DRYING WET	Method and apparat	🌐	A1	WO-2005123885-A
□	0.00(		33552008	🇬🇧	BIOFUEL CELL	The present inventi	🌐	A2	WO-2005001981-A
□	0.00(		34192402	🇬🇧	Method for manufacturing bio-diesel oil containing	A method for manuf	🇺🇸	A1	US-20050039383-A
□	0.00(		34972849	🇬🇧	METHOD AND APPARATUS FOR DRYING WET	Method and apparat	🇨🇦	A1	CA-2570278-A1
□	0.00(		32906804	🌐	A FLOWABLE BIO-FUEL MATERIAL FROM CERE	The present inventi	🌐	A1	WO-2005123886-A
□	0.00(		34827329	🇬🇧	Thermobaric molecular fractionation	Systems and meth	🇺🇸	A1	US-20050177013-A
□	0.00(		34739126	🇬🇧	Integrated process for producing _clean beef_ (or	An ethanol unit and	🇺🇸	A1	US-20050153410-A
□	0.00(		35506189	🇬🇧	Bio-battery	A bio-battery include	🇺🇸	A1	US-20050287399-A
□	0.00(		34966649	🇬🇧	METHOD FOR THE PRODUCTION OF NATURAI	The method aims a	🌐	A1	WO-2005102547-A
□	0.00(		34042906	🇬🇧	BIO-ELECTROCHEMICAL PROCESS FOR PRO	A process for produ	🌐	A2	WO-2005005981-A
□	0.00(		35134425	🇬🇧	Biomass converters and processes	This invention perta	🇺🇸	A1	US-20050217134-A
□	0.00(		32500562	🇬🇧	Use of a mixture of esters of fatty acids as fuel or	The invention relate	🇺🇸	A1	US-20050261144-A
□	0.00(		32500562	🇬🇧	USE OF A MIXTURE OF ESTERS OF FATTY ACI	The invention relate	🇪🇺	A1	EP-1569923-A1
□	0.00(		34740776	🇬🇧	Process for the preparation of bio-diesel	A process is disclos	🇺🇸	A1	US-20050108927-A
□	0.00(		35479149	🇬🇧	Hybrid generation with alternative fuel sources	A generating facility	🇺🇸	A1	US-20050279095-A
□	0.00(		34836933	🇬🇧	PROCESS FOR MAKING SOLID FUEL	This invention provi	🌐	A1	WO-2005075609-A

Facets

Kind Codes of Documents

- A1 (47267)
- A (17800)
- A2 (10038)
- B2 (8663)
- B1 (6561)
- A3 (4190)
- C (4101)
- A4 (1458)
- B (550)
- A8 (407)

more...

Publication date of documents

- 2009 (14822)
- 2008 (13881)
- 2007 (13093)
- 2006 (11525)
- 2005 (10591)
- 2004 (9995)

Exploratory Search

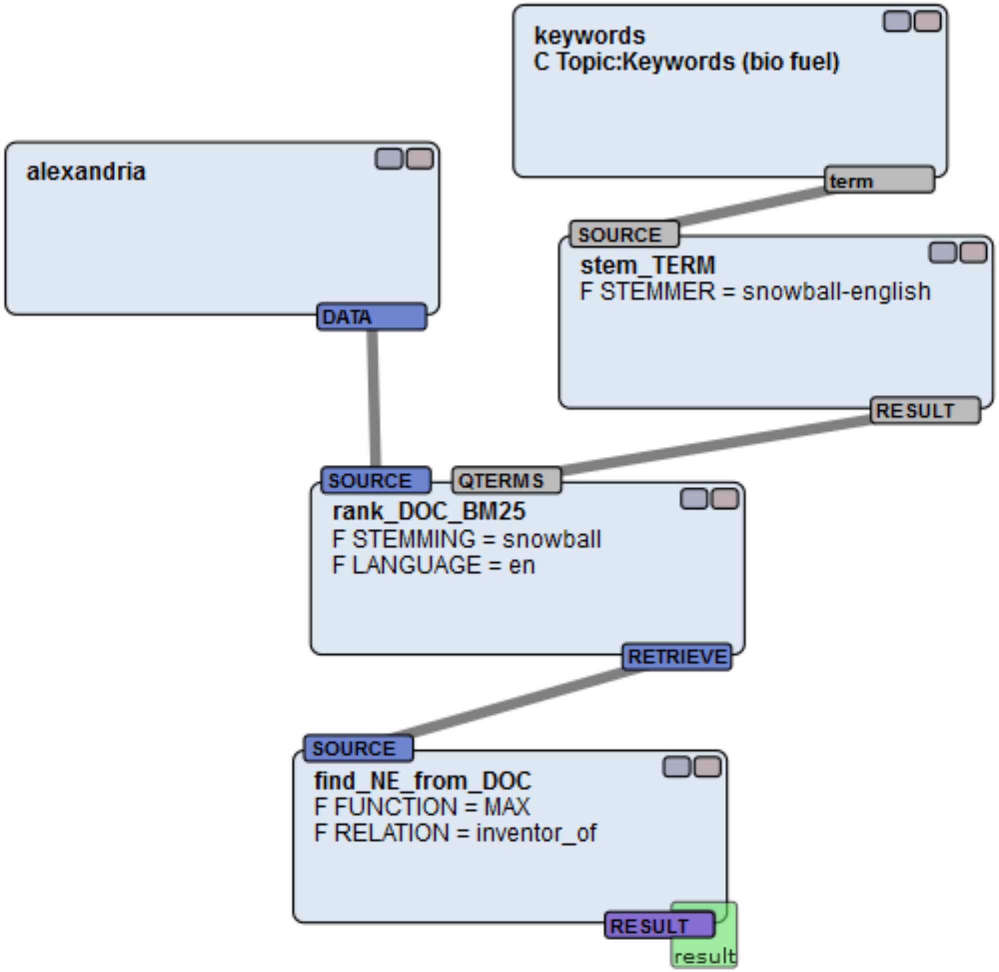
- Search & (Faceted) Browsing
  - Help discover schema, ontology, etc.
  - Help discover the relevant sources
    - Within-collection (by year/location, by type, …)
    - Across multiple collections (by source)

  - *Tony Russel Rose is likely to tell us more later this afternoon!*

Exploratory Search

- PRA enables soft (or “fuzzy”) faceted selections
  - Re-weight based on preferences, no more zero-result-set problem!

From Patent to Inventor



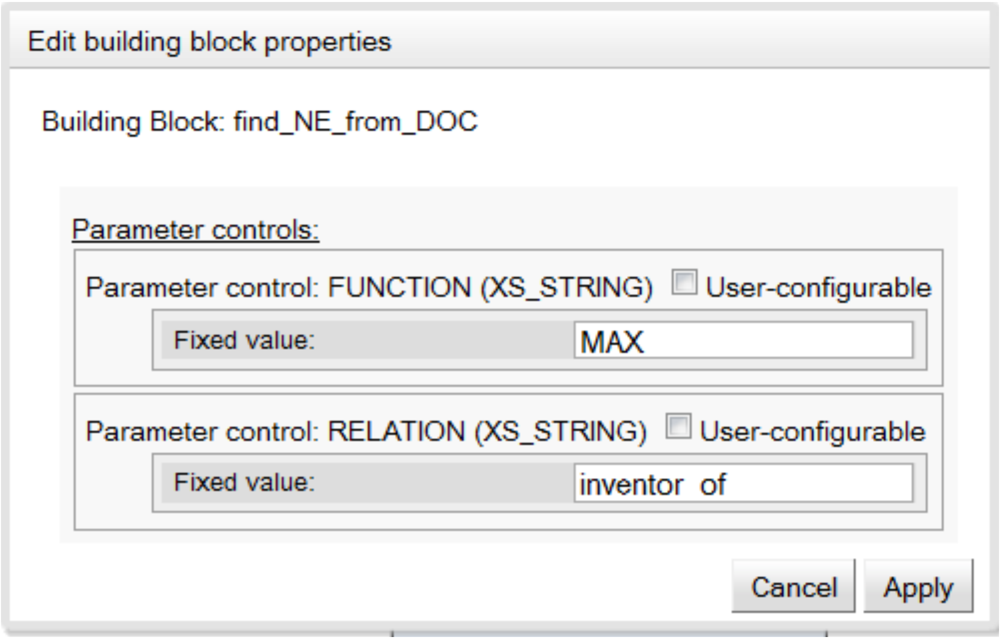
keywords
C Topic:Keywords (bio fuel)

alexandria

stem_TERM
F STEMMER = snowball-english

rank_DOC_BM25
F STEMMING = snowball
F LANGUAGE = en

find_NE_from_DOC
F FUNCTION = MAX
F RELATION = inventor_of



Edit building block properties

Building Block: find_NE_from_DOC

Parameter controls:

Parameter control: FUNCTION (XS_STRING) □ User-configurable

Fixed value: MAX

Parameter control: RELATION (XS_STRING) □ User-configurable

Fixed value: inventor_of

Limitations Search & Browse

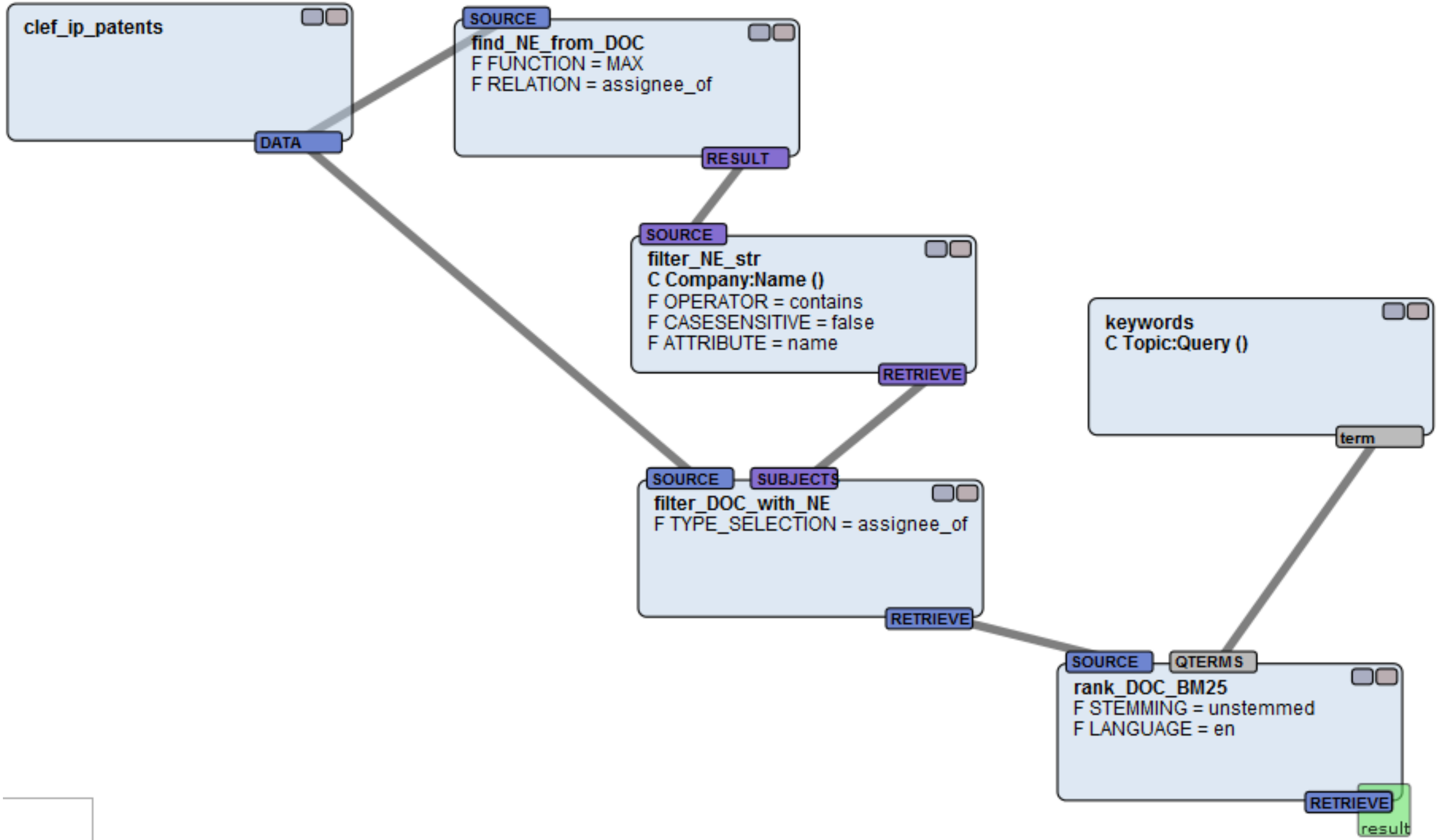
- Faceted exploration does not include joins
  - Cannot construct new data sources from existing ones!
  - Only the pre-defined paths through the information space can actually be traversed

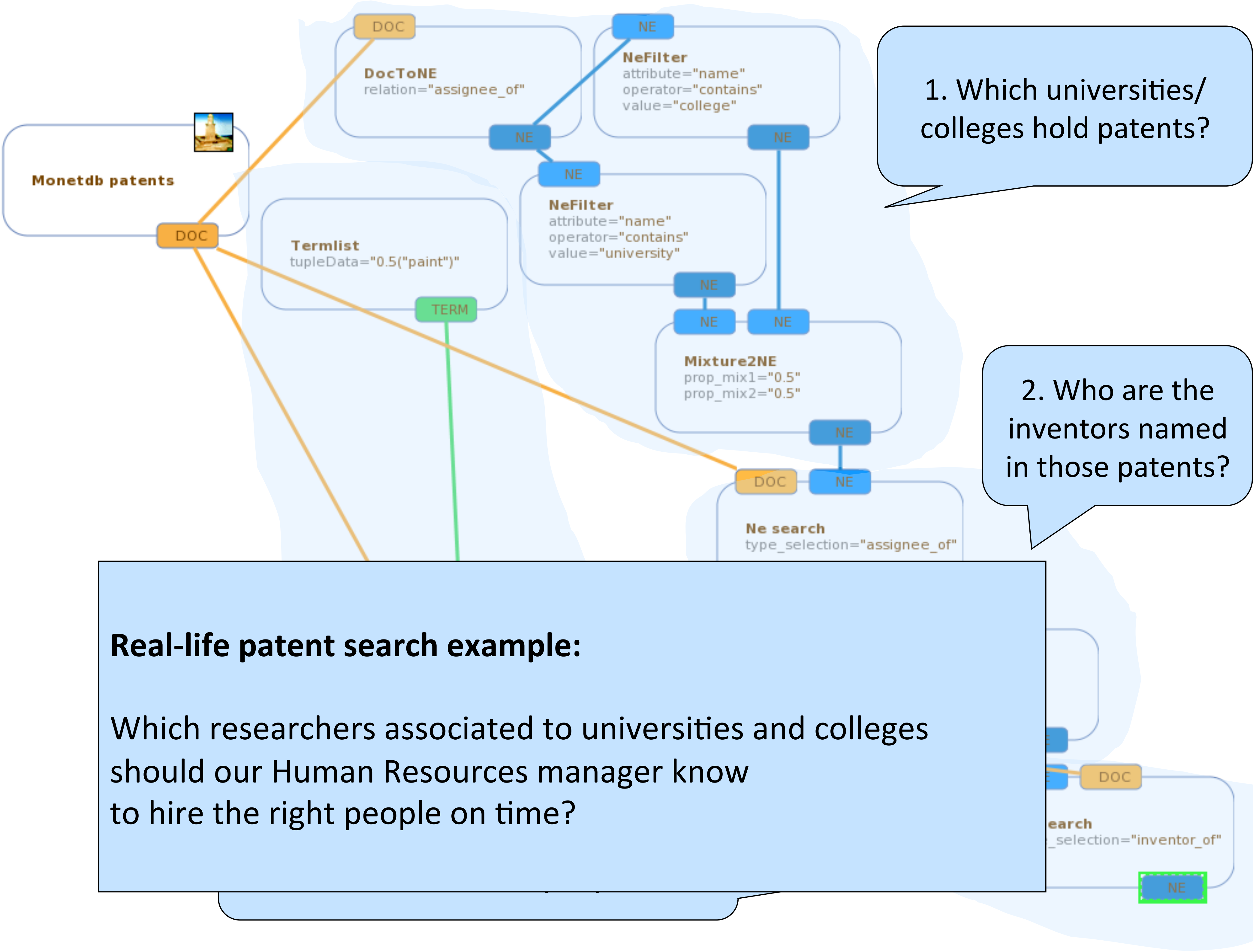
Who needs a Join?

- You!!!
  … whenever ‘relevance cues’ are typed:
  - People (e.g., inventors)
  - Companies (e.g., assignees)
  - Categories (e.g., IPTC)
  - Time (e.g., expiry date)
  - Location (e.g., country)

  … or whenever multiple sources are to be combined
  - E.g., patents & news, patents & Wikipedia, …

Patents on X by Y(y)





Monetdb patents

DocToNE
relation="assignee_of"

NeFilter
attribute="name"
operator="contains"
value="college"

NeFilter
attribute="name"
operator="contains"
value="university"

Termlist
tupleData="0.5("paint")"

Mixture2NE
prop_mix1="0.5"
prop_mix2="0.5"

Ne search
type_selection="assignee_of"

earch
_selection="inventor_of"

1. Which universities/ colleges hold patents?

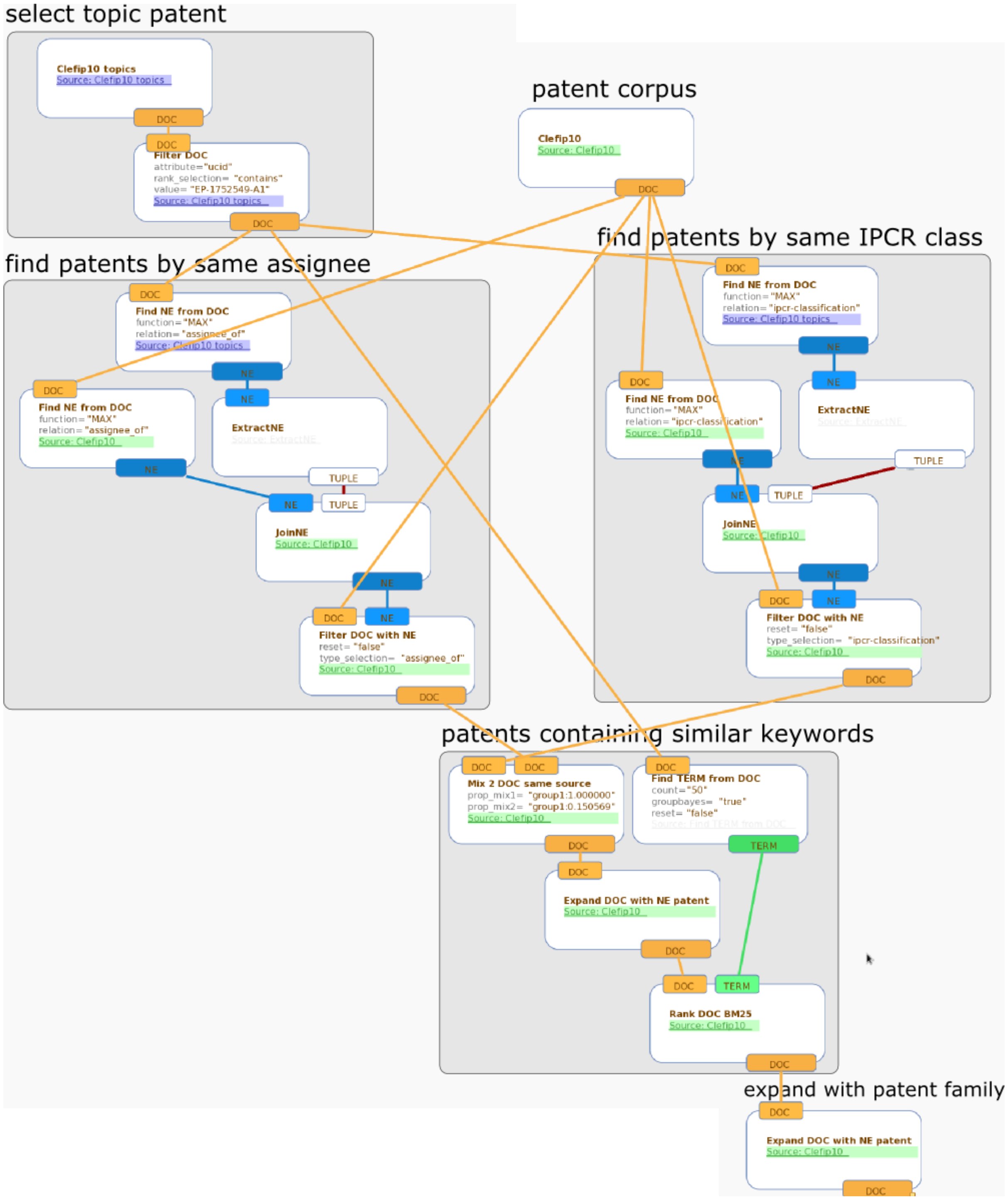
2. Who are the inventors named in those patents?

Real-life patent search example:

Which researchers associated to universities and colleges
should our Human Resources manager know
to hire the right people on time?

How Strategies Help

- Strategies improve communication between search intermediary and user
  - Encapsulate domain expert knowledge
  - Abstract representation of search expert knowledge
  - Analyze information seeking process at any stage
- Strategies facilitate knowledge management
  - Store / share / publish / refine
- Strategies mix exact (DB) and ranked (IR) searches
  - Avoid the need for "human (probabilistic) joins"


select topic patent
patent corpus
find patents by same assignee
find patents by same IPCR class
patents containing similar keywords
expand with patent family
DOC
NE
TUPLE
TERM

Clefip10 topics
Source: Clefip10 topics

Filter DOC
attribute="ucid"
rank_selection= "contains"
value= "EP-1752549-A1"
Source: Clefip10 topics

Clefip10
Source: Clefip10

Find NE from DOC
function="MAX"
relation="assignee_of"
Source: Clefip10 topics

Find NE from DOC
function="MAX"
relation="assignee_of"
Source: Clefip10

ExtractNE
Source: ExtractNE

JoinNE
Source: Clefip10

Filter DOC with NE
reset= "false"
type_selection= "assignee_of"
Source: Clefip10

Find NE from DOC
function="MAX"
relation="ipcr-classification"
Source: Clefip10 topics

Find NE from DOC
function="MAX"
relation="ipcr-classification"
Source: Clefip10

ExtractNE
Source: ExtractNE

JoinNE
Source: Clefip10

Filter DOC with NE
reset= "false"
type_selection= "ipcr-classification"
Source: Clefip10

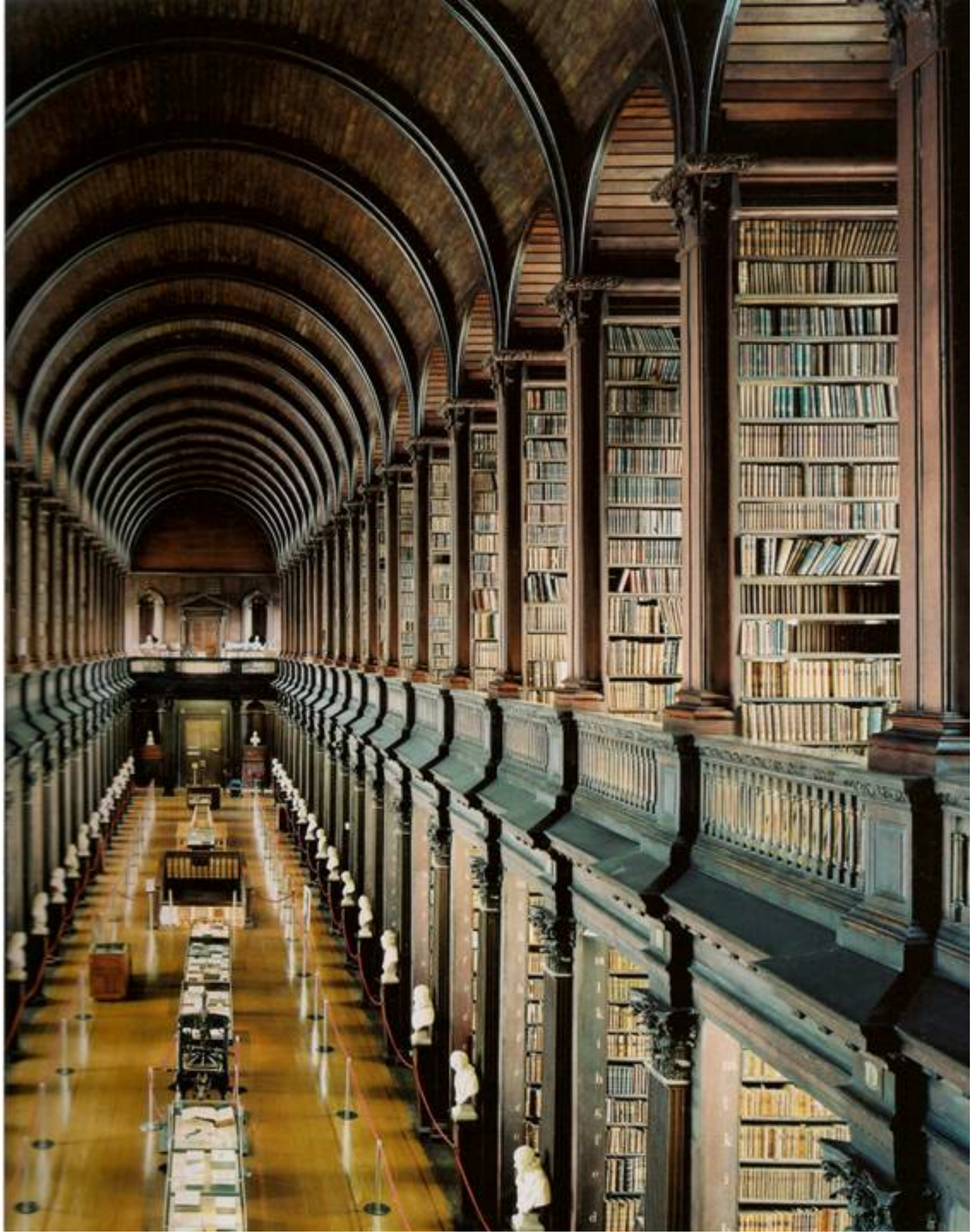
Mix 2 DOC same source
prop_mix1= "group1:1.000000"
prop_mix2= "group1:0.150569"
Source: Clefip10

Find TERM from DOC
count="50"
groupbayes= "true"
reset= "false"
Source: Find TERM from DOC

Expand DOC with NE patent
Source: Clefip10

Rank DOC BM25
Source: Clefip10

Expand DOC with NE patent
Source: Clefip10



Trinity College Library, Dublin

Conclusion

- “No idealized one-shot search engine”
- Hand over control to the user (or, most likely, the search intermediary)
  - Patent information specialists
  - Digital forensics detectives
  - Librarians / archivists
  - Real estate agents
  - Travel agency

Interactive Information Access

- *Feedback:*
  - Interaction improves information representation
- *Faceted Browsing:*
  - Interaction can let user take over where machine would fail
- *Search by Strategy:*
  - Interaction can let user take over where system designer would fail

Research Opportunities

- Assist the user make the best out of their increased level of control
  - Integrate usage data from live system to help improve or adapt strategies
- Handle “even larger” scale data
  - Patent demo fine on ~17GB semi-structured data (i.e., Fairview Research’s Green Energy collection), without specific optimizations, even with fairly large strategies
- Formalism
  - Score normalization
- Close the loop!

Current Situation

- index ;
- repeat {
-         specify ;
-         retrieve
- } until ☺

Schema definition

Search & explore

Desirable Situation

```
repeat {
        index ;
        specify ;
        retrieve
} until ☺
```

Mixed Initiative
Schema definition
Search & explore

Österreichische Nationalbibliothek

Acknowledgements

Wouter Alink	Spinque BV
Roberto Cornacchia	(co-founders)
Martin Kersten & team	CWI/MonetDB BV
Thomas Rölleke	QMUL/A Priori
Henk Tomas	IP specialist
Francisco Webber	IRF & Matrixware