MicroarrayWEB: A WEB application for microarray data handling

Antonio d'Acierno
ISA – CNR
antonio.dacierno@gmail.com
http://www.isa.cnr.it/dacierno/

Objectives

- To realize a WEB application mainly able to
  - Store and retrieve experimental data
  - Searching experiments among stored data
    - Using meta data
    - Using experimental data

Problem: How to describe an experiment?

- MIAME
  - **Minimum Information About a Microarray Experiment**
  - MIAME describes the information about a microarray experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment.

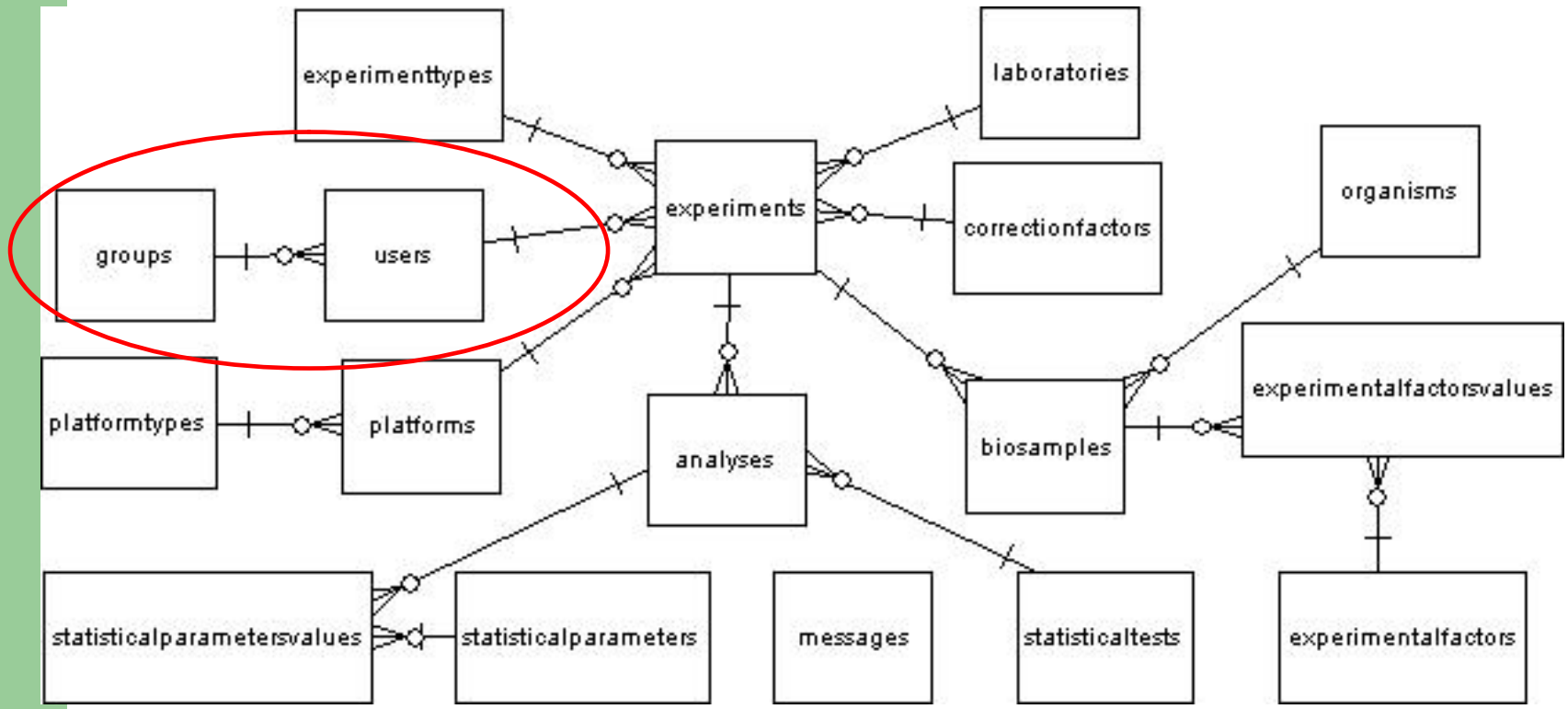
MIAME (1/2)

- The raw data for each hybridization (e.g., CEL or GPR files)
- The final processed (normalized) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
- The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)

MIAME (2/2)

- The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates)
- Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

MIAME as an ER Model



Data Privacy

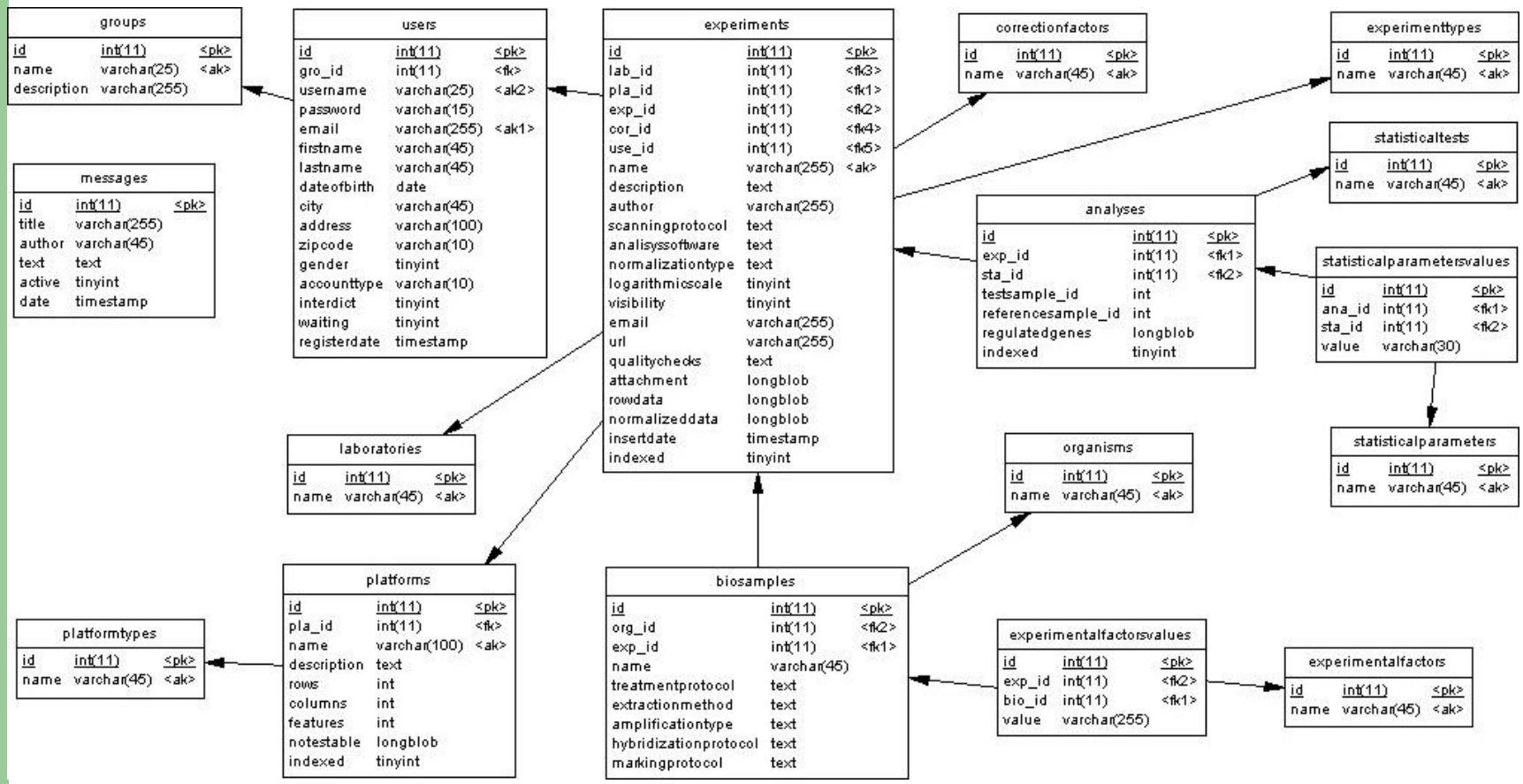
- Users are organized in
  - users
  - registered users
  - administrators
- Registered users are grouped into *Research Teams*
- Experiments are labeled as
  - Public, Protected and Private

The Experiment Entity
The Analysis Entity

experiments			
id	<pi>	Serial (11)	<M>
name	<ai>	Variable characters (255)	<M>
description		Text	
author		Variable characters (255)	<M>
scanningprotocol		Text	
analisyssoftware		Text	
normalizationtype		Text	
logarithmicscale		Byte (1)	
visibility		Byte (4)	<M>
email		Variable characters (255)	
url		Variable characters (255)	
qualitychecks		Text	
attachment		Long binary	
rowdata		Long binary	
normalizeddata		Long binary	
insertdate		Timestamp	<M>
indexed		Byte (1)	<M>

analyses			
id	<pi>	Serial (11)	<M>
testsample_id		Integer	<M>
referencesample_id		Integer	<M>
regulatedgenes		Long binary	
indexed		Byte (1)	<M>

MIAME as a logical Model



Using a RDB based on MIAME ….

- We can develop a (web based) application to store and retrieve microarray data.
- We can protect data using users, groups and visibility levels (public, protected, private).
- We can navigate our data using MIAME metadata.
- Our system does not add value to, for example, BASE2.
- To add value we have to implement advanced search capabilities
  - i.e., for example, find the experiments where a given gene has been analyzed

Notes table example

TargetID	ProbeID	ONTOLOGY_COMPONENT
15E1.2	20605	cellular component unknown [goid 8372] [evidence ND]
ABCA2	3520743	integral to membrane [goid 16021] [evidence NAS]; ATF-binding cassette (ABC) transporter complex [goid 43190] [pmid 11178988] [evidence NAS]; lysosomal membrane [goid 5765] [pmid 11309290] [evidence IDA]
76P	3060450	centrosome [goid 5813] [pmid 10562286] [evidence TAS]; microtubule [goid 5874] [evidence IEA]; gamma-tubulin ring complex [goid 8274] [pmid 10562286] [evidence NAS]
AADAC	580711	endoplasmic reticulum membrane [goid 5789] [pmid 15152005] [evidence IDA]; membrane [goid 16020] [evidence IEA]; microsome [goid 5792] [pmid 8063807] [evidence TAS]; integral to membrane [goid 16021] [evidence IEA]
A4CNT	4590047	membrane fraction [goid 5624] [pmid 10430883] [evidence TAS]; integral to membrane [goid 16021] [pmid 10430883] [evidence TAS]; Golgi stack [goid 5795] [evidence IEA]
A2BP1	770300	Golgi apparatus [goid 5794] [pmid 10814712] [evidence TAS]
A2BP1	3290546	Golgi apparatus [goid 5794] [pmid 10814712] [evidence TAS]
A2BP1	3420601	Golgi apparatus [goid 5794] [pmid 10814712] [evidence TAS]
A2M	7400044	extracellular region [goid 5576] [pmid 14718574] [evidence NAS]
A4GALT	5670674	membrane fraction [goid 5624] [pmid 10748143] [evidence IDA]; integral to Golgi membrane [goid 30173] [pmid 10748143] [evidence NAS]; integral to membrane [goid 16021] [evidence IEA]; Golgi stack [goid 5795] [evidence IEA]; membrane [goid 16020] [evidence IEA]

Normalized data example

ProbeId	Symbol	O_0_17_35_B	O_0_17_51_B	O_0_17_52_A	O_3_17_35_D	O_3_17_51_D	O_3_17_52_C	O_6_17_56_C	O_6_17_57_C	O_6_17_57_G
1450041	**Dclre1c**	181	189	159	179	177	188	156	158	160
4540037	**Mdm1**	202	165	196	208	221	214	180	210	214
1780056	**Zp3r**	139	187	187	177	196	187	183	172	166
610408	**E430018M08Rik**	185	168	172	165	150	136	198	162	181
610369	**Ampd3**	184	181	143	163	153	144	199	147	198
1450014	**D830016O14Rik**	141	107	130	149	136	158	147	165	145
380019	**Crry**	318	342	309	308	347	334	431	479	318
6860707	**4930470D19Rik**	234	247	291	217	194	258	219	224	237
1850619	**Sfrs16**	148	113	155	162	144	141	134	147	149
3780279	**Prkcz**	681	702	640	428	449	433	339	362	333
610088	**2810028N01Rik**	159	165	190	160	155	181	179	194	199

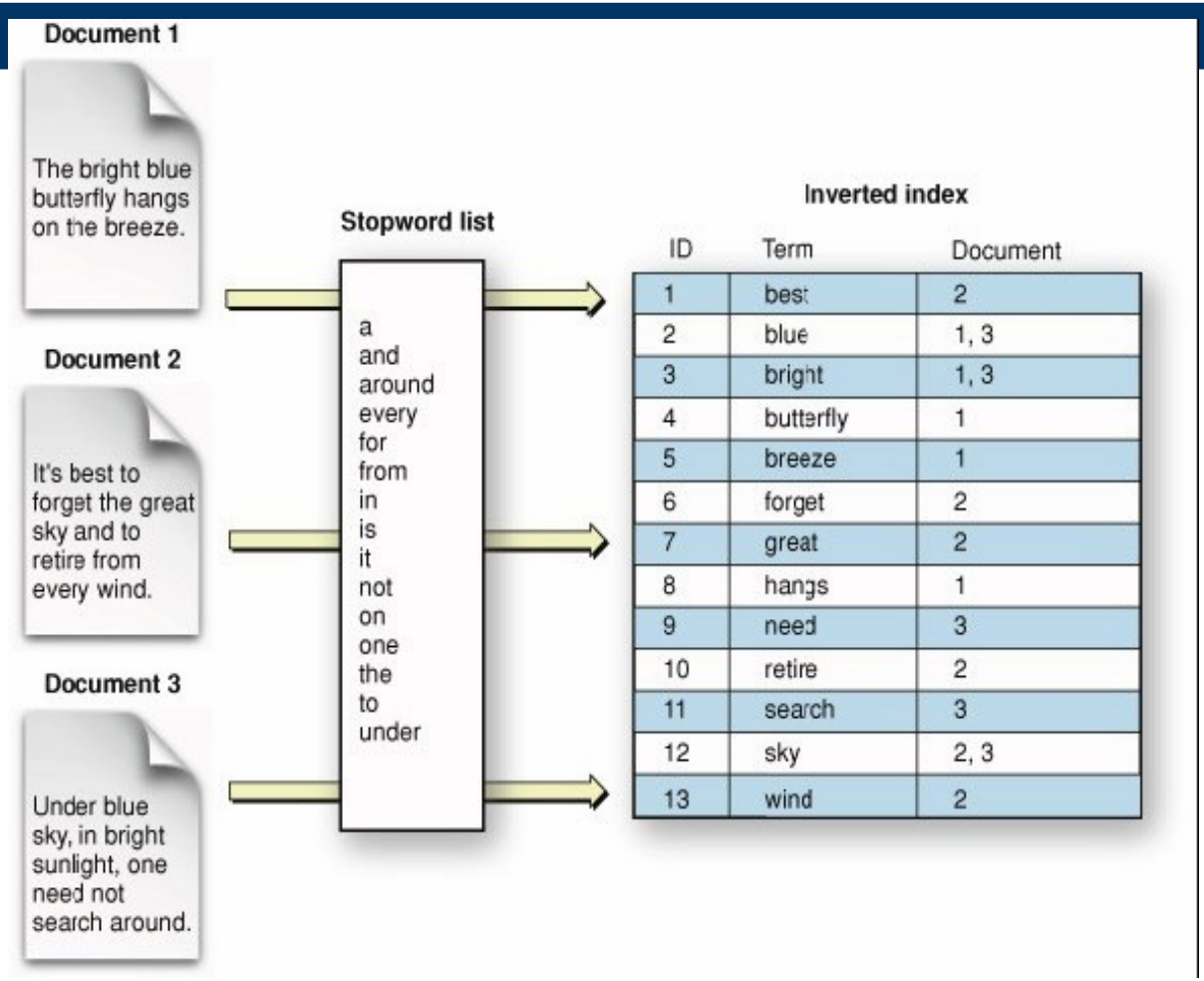
Are IR techniques applicable?

- For each experiment we have to store:
  - Raw Data
    - a compressed file (stored as a blob)
  - Normalized Data
    - a text file (stored as a blob)
  - Lists of Regulated Genes
    - a text file for each analysis of the experiment (stored as a blob)
- For each platform
  - The notes table (stored as a blob)
    - From this table it is possible have information about probe and gene(s).
- We do not use MySQL full-text search capabilities

Apache Lucene

- Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java.
- Implementations in other programming languages are available (index-compatible)
- It is an open source technology suitable for nearly any application that requires full-text search

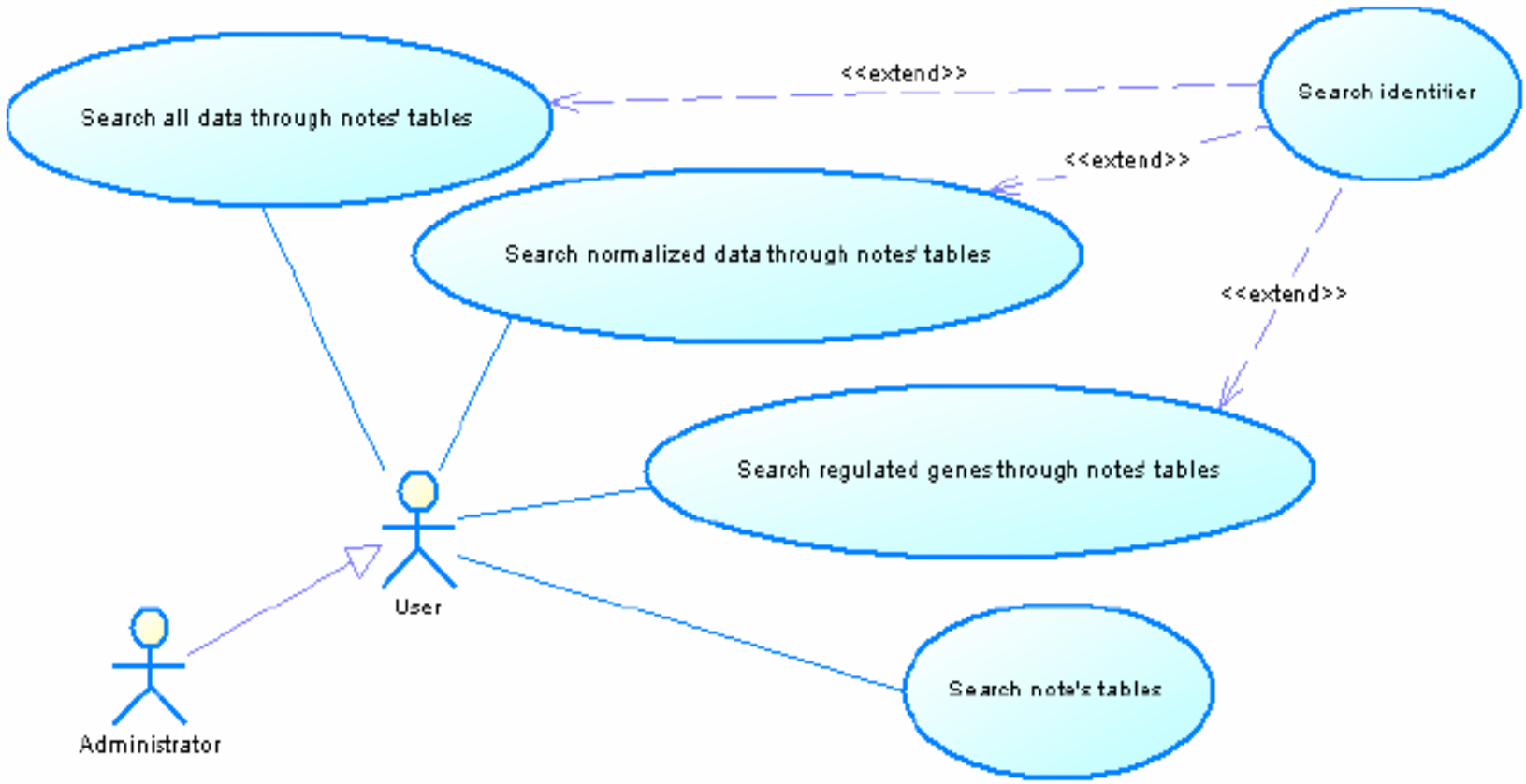
The Lucene Inverted Index



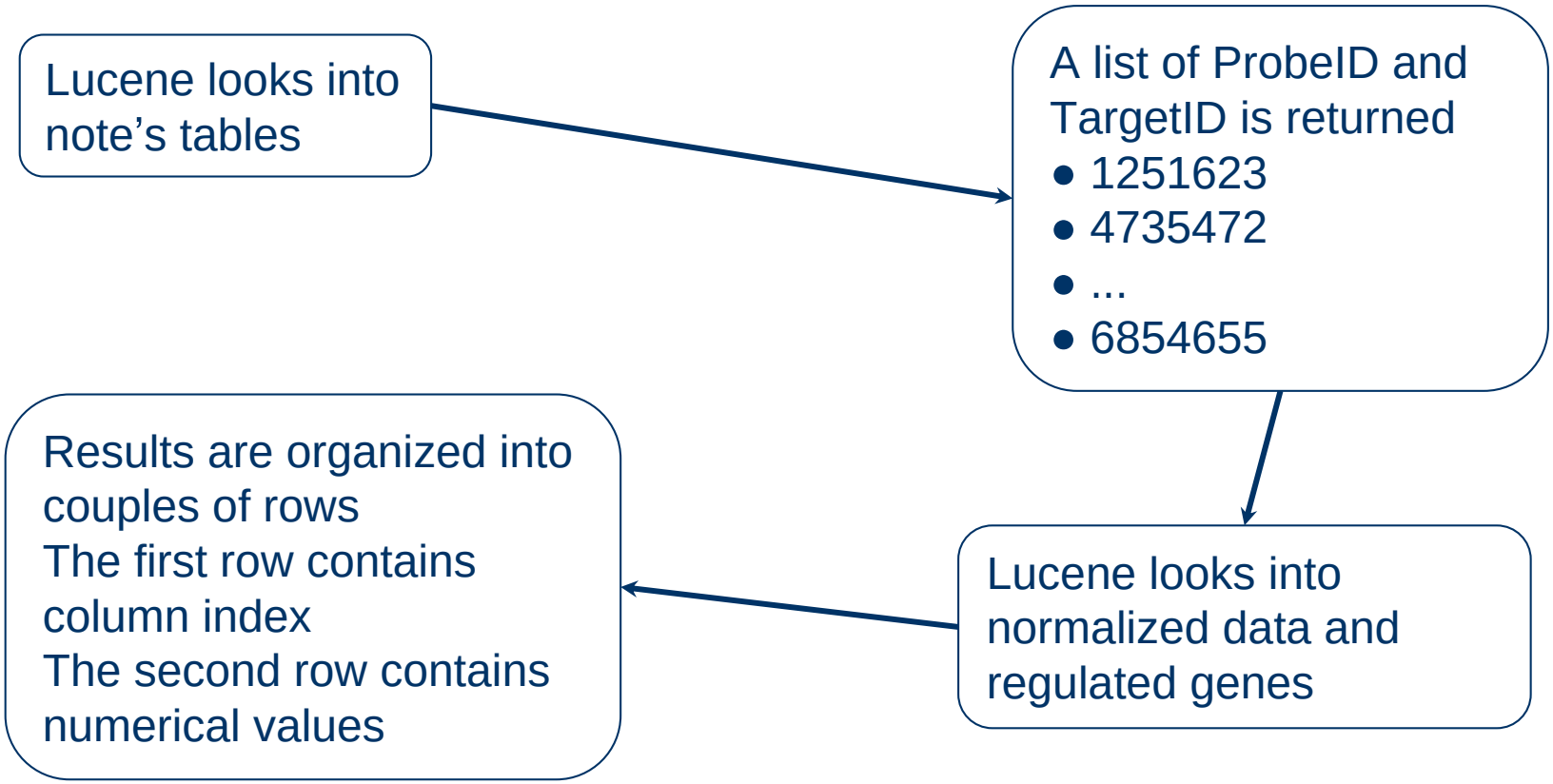
Apache Lucene advantages

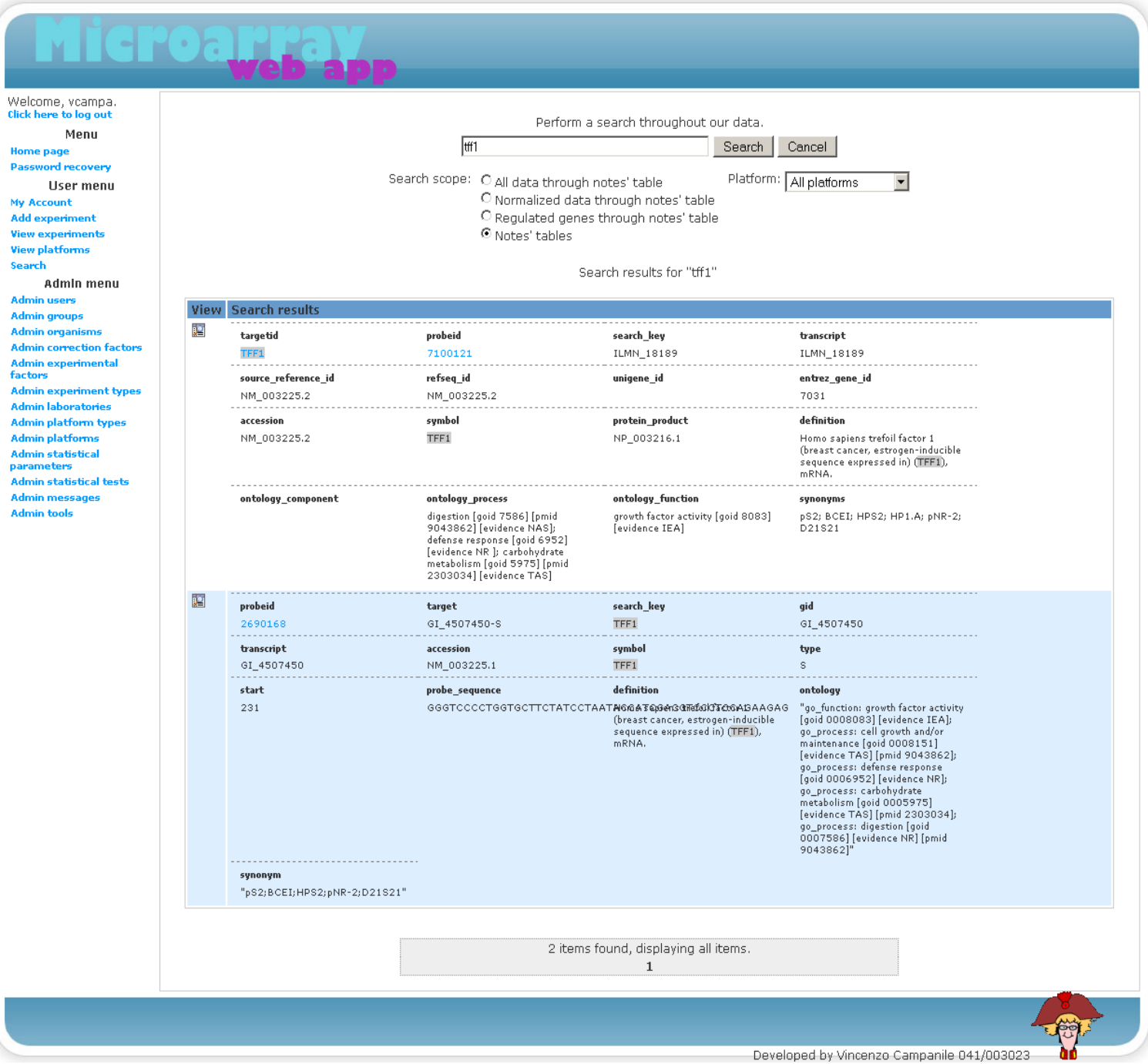
- not only text files
- stop words support
- stemming
- concurrent use of the index
- multi language support
- to filter research output
- automatic ordering according to a score
- queries with wild cards
- range queries and natural language queries
- proximity queries and fuzzy queries
- queries caching
- term boosting
- highlighting support
- batch indexing

Information Retrieval in MicroarrayWEB



Example:
Search all data through note's tables




Microarray
web app
Welcome, vcampa.
Click here to log out
Menu
Home page
Password recovery
User menu
My Account
Add experiment
View experiments
View platforms
Search
Admin menu
Admin users
Admin groups
Admin organisms
Admin correction factors
Admin experimental factors
Admin experiment types
Admin laboratories
Admin platform types
Admin platforms
Admin statistical parameters
Admin statistical tests
Admin messages
Admin tools
Perform a search throughout our data.
tff1
Search
Cancel
Search scope:
All data through notes' table
Normalized data through notes' table
Regulated genes through notes' table
Notes' tables
Platform:
All platforms
Search results for "tff1"
View
Search results
targetid
TFF1
probeid
7100121
search_key
ILMN_18189
transcript
ILMN_18189
source_reference_id
NM_003225.2
refseq_id
NM_003225.2
unigene_id
entrez_gene_id
7031
accession
NM_003225.2
symbol
TFF1
protein_product
NP_003216.1
definition
Homo sapiens trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in) (TFF1), mRNA.
ontology_component
ontology_process
digestion [goid 7586] [pmid 9043862] [evidence NAS]; defense response [goid 6952] [evidence NR]; carbohydrate metabolism [goid 5975] [pmid 2303034] [evidence TAS]
ontology_function
growth factor activity [goid 8083] [evidence IEA]
synonyms
pS2; BCEI; HPS2; HP1.A; pNR-2; D21S21
probeid
2690168
target
GI_4507450-S
search_key
TFF1
gid
GI_4507450
transcript
GI_4507450
accession
NM_003225.1
symbol
TFF1
type
S
start
231
probe_sequence
definition
(breast cancer, estrogen-inducible sequence expressed in) (TFF1), mRNA.
ontology
"go_function: growth factor activity [goid 0008083] [evidence IEA]; go_process: cell growth and/or maintenance [goid 0008151] [evidence TAS] [pmid 9043862]; go_process: defense response [goid 0006952] [evidence NR]; go_process: carbohydrate metabolism [goid 0005975] [evidence TAS] [pmid 2303034]; go_process: digestion [goid 0007586] [evidence NR] [pmid 9043862]"
synonym
"pS2;BCEI;HPS2;pNR-2;D21S21"
2 items found, displaying all items.
1
Developed by Vincenzo Campanile 041/003023

Microarray web app

Welcome, vcampa.
Click here to log out

Menu

- Home page
- Password recovery

User menu

- My Account
- Add experiment
- View experiments
- View platforms
- Search

Admin menu

- Admin users
- Admin groups
- Admin organisms
- Admin correction factors
- Admin experimental factors
- Admin experiment types
- Admin laboratories
- Admin platform types
- Admin platforms
- Admin statistical parameters
- Admin statistical tests
- Admin messages
- Admin tools

Perform a search throughout our data.

sodium | Search | Cancel

Search scope:
- (•) All data through notes' table
- ( ) Normalized data through notes' table
- ( ) Regulated genes through notes' table
- ( ) Notes' tables

Platform: All platforms

Search results for "sodium"

View	Search results			
	probeid	str_0_17-65:avg_signal	mcf-7_nt1_17-65:avg_signal	mcf-7_t1_17-65:avg_signal
	2140446	-34.61121	-40.98322	4.541316
	mcf-7_nt2_17-65:avg_signal	mcf-7_t2_17-65:avg_signal	a-549_nt1_17-65:avg_signal	a-549_t1_17-66:avg_signal
	-17.87215	15.82704	11.04752	4.385753
	a-549_nt2_17-66:avg_signal	a-549_t2_17-66:avg_signal	mcf-7_nt1_17-66:avg_signal	mcf-7_t1_17-66:avg_signal
	-17.51368	4.638127	-40.98322	4.541316
	mcf-7_nt2_17-66:avg_signal	mcf-7_t2_17-67:avg_signal	a-549_nt1_17-67:avg_signal	a-549_t1_17-67:avg_signal
	-17.87215	15.82704	11.04752	4.385753
	a-549_nt2_17-67:avg_signal	a-549_t2_17-67:avg_signal	zr-75.1_0_17-67:avg_signal	
	-17.51368	4.638127	34.59727	
	probeid	str_0_17-65:avg_signal	mcf-7_nt1_17-65:avg_signal	mcf-7_t1_17-65:avg_signal
	2140446	-34.61121	-40.98322	4.541316
	mcf-7_nt2_17-65:avg_signal	mcf-7_t2_17-65:avg_signal	a-549_nt1_17-65:avg_signal	a-549_t1_17-66:avg_signal
	-17.87215	15.82704	11.04752	4.385753
	a-549_nt2_17-66:avg_signal	a-549_t2_17-66:avg_signal	mcf-7_nt1_17-66:avg_signal	mcf-7_t1_17-66:avg_signal
	-17.51368	4.638127	-40.98322	4.541316
	mcf-7_nt2_17-66:avg_signal	mcf-7_t2_17-67:avg_signal	a-549_nt1_17-67:avg_signal	a-549_t1_17-67:avg_signal
	-17.87215	15.82704	11.04752	4.385753
	a-549_nt2_17-67:avg_signal	a-549_t2_17-67:avg_signal	zr-75.1_0_17-67:avg_signal	
	-17.51368	4.638127	34.59727	
	probeid	str_0_17-65:avg_signal	mcf-7_nt1_17-65:avg_signal	mcf-7_t1_17-65:avg_signal
	2140446	-34.61121	-40.98322	4.541316
	mcf-7_nt2_17-65:avg_signal	mcf-7_t2_17-65:avg_signal	a-549_nt1_17-65:avg_signal	a-549_t1_17-66:avg_signal
	-17.87215	15.82704	11.04752	4.385753
	a-549_nt2_17-66:avg_signal	a-549_t2_17-66:avg_signal	mcf-7_nt1_17-66:avg_signal	mcf-7_t1_17-66:avg_signal

Open Problems

- We would like to test the application.
- Redundancy: we store some data twice; is this a problem?
- A very high number of documents.
- The possibility of using a dedicated indexing server should be investigated.