The LAILAPS plant science search engine: Explore plant genome databases

Jinbo Chen, Christian Colmsee, Uwe Scholz and Matthias Lange
(Research Group BIT, IPK Gatersleben, Germany)
14 October 2015 – NETTAB & IB 2015

Outline

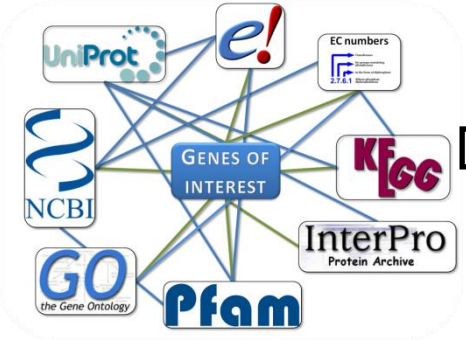
- Motivation
- Overview
- Implementation and Example
- Technology Details

Search Engines in Life Science

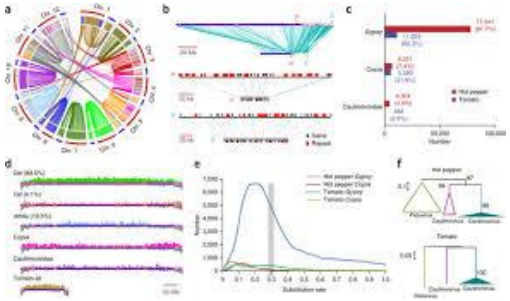
Information search for for Hordeum vulgare (barley)

Information search gets an difficult and time consuming task in a heterogeneous ecosystem of life science

LAILAPS Integrated Search



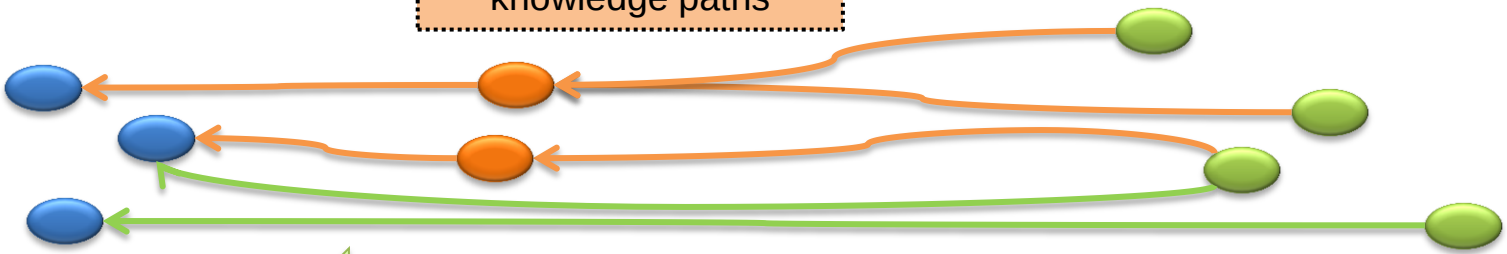
LAILAPS
THE PLANT SCIENCE SEARCH ENGINE



textual notated knowledge
- literature
- ontologies
- curated literature excerpts (proteins, pathways)

none textual facts databases
- genome
- metabolome
- phenome

indirect / transitive knowledge paths

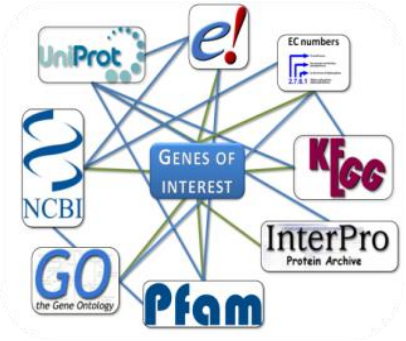


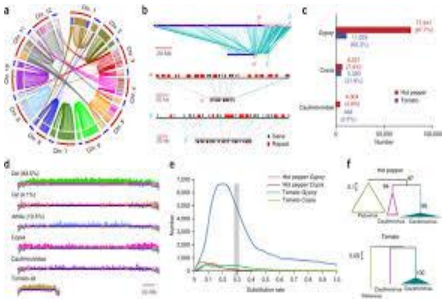
direct functional annotation
- homologies, wet lab, …

Information Retrieval Components

1. **Index**
   - text / data decomposition
   - language processing
   - synonyms, homonyms
2. **Query**
   - efficient search in content
3. **Ranking**
   - feature extraction
   - ranking functions
   - pertinence (subjective user relevance profiles)
4. **Presentation**
   - intuitive user interface
   - related entries („page like this“)
   - query suggestion („did you mean“)

Indexing: Running Instances*



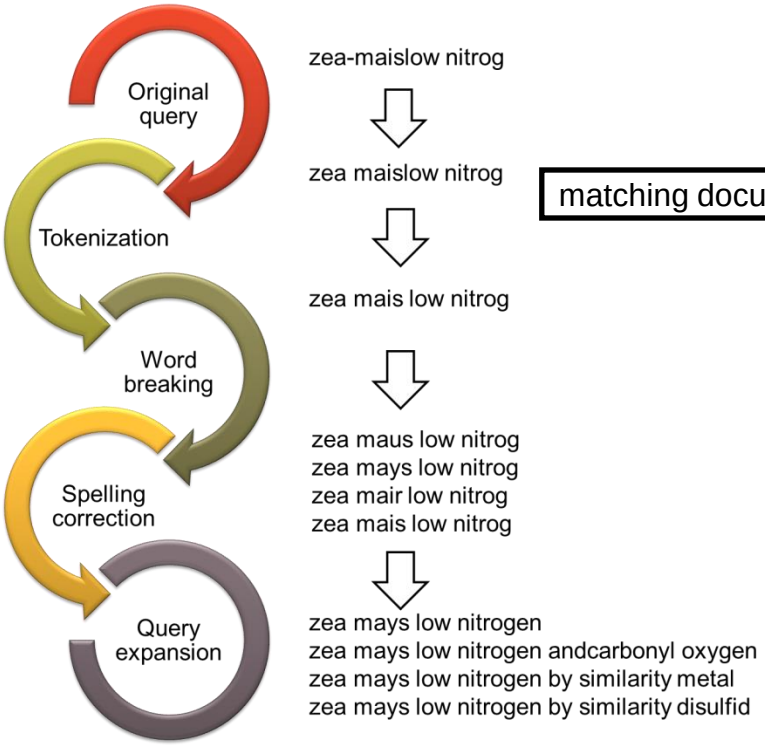


		indexed documents	linked records
transPlant, EU/UK	gene models, protein, ontologies, literature	$63.5*10^6$	$50*10^6$
IPK, Germany	plant genomes, germplasm collections	$62.7*10^6$	$3*10^6$
Helmholz Center Munich, Germany	barley and wheat genome, literature	$60.5*10^6$	$0.6*10^6$
Barcelona Supercomp. Center, Spain	IPK mirror	$63.5*10^6$	$50*10^6$

*http://lailaps.ipk-gatersleben.de/products/running_instance.html

Query and Ranking

query workflow

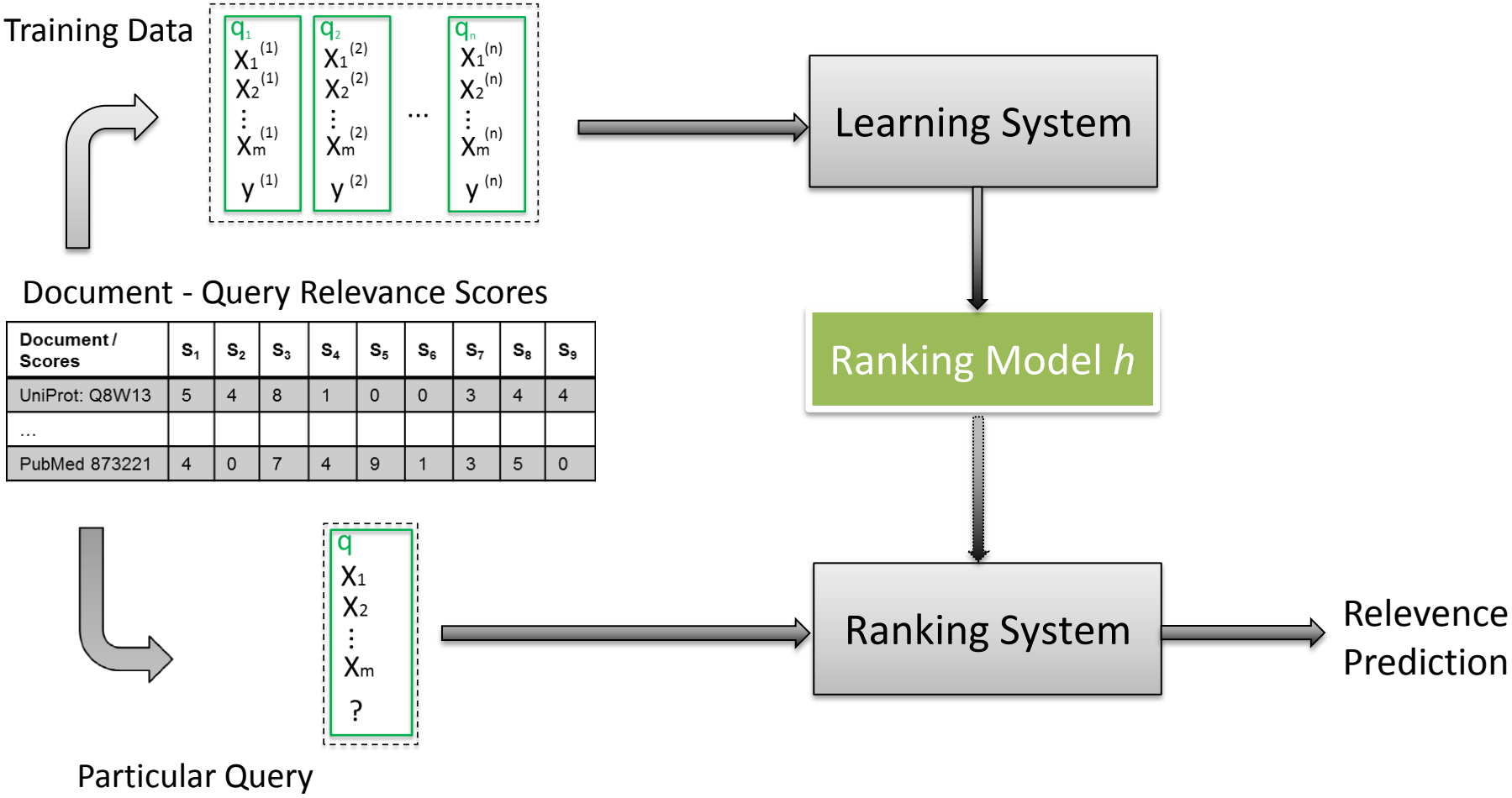


matching documents

query result

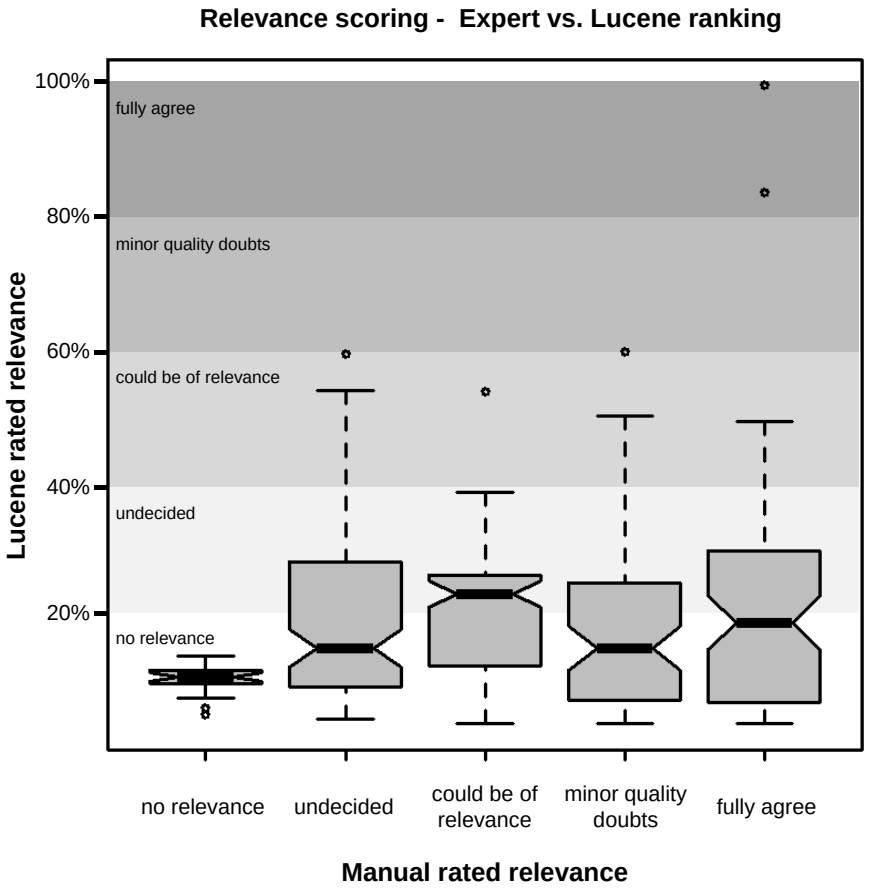
Document / Scores	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
UniProt: Q8W13	5	4	8	1	0	0	3	4	4
…									
PubMed 873221	4	0	7	4	9	1	3	5	0

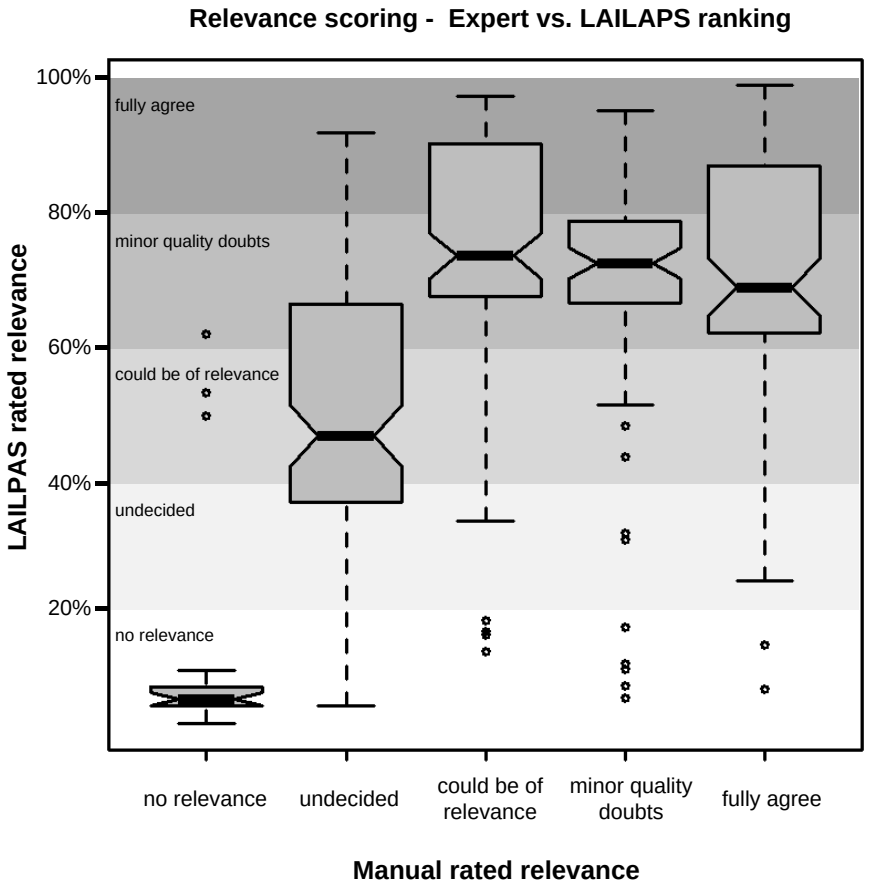
Query and Ranking



Document / Scores	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
UniProt: Q8W13	5	4	8	1	0	0	3	4	4
...									
PubMed 873221	4	0	7	4	9	1	3	5	0

LAILAPS Ranking Results





Esch et al. (*Plant and Cell Physiology*, Database Issue 2015)

LAILAPS Example

Query:
"salt stres barley"

Spelling correction & synonym expansion:

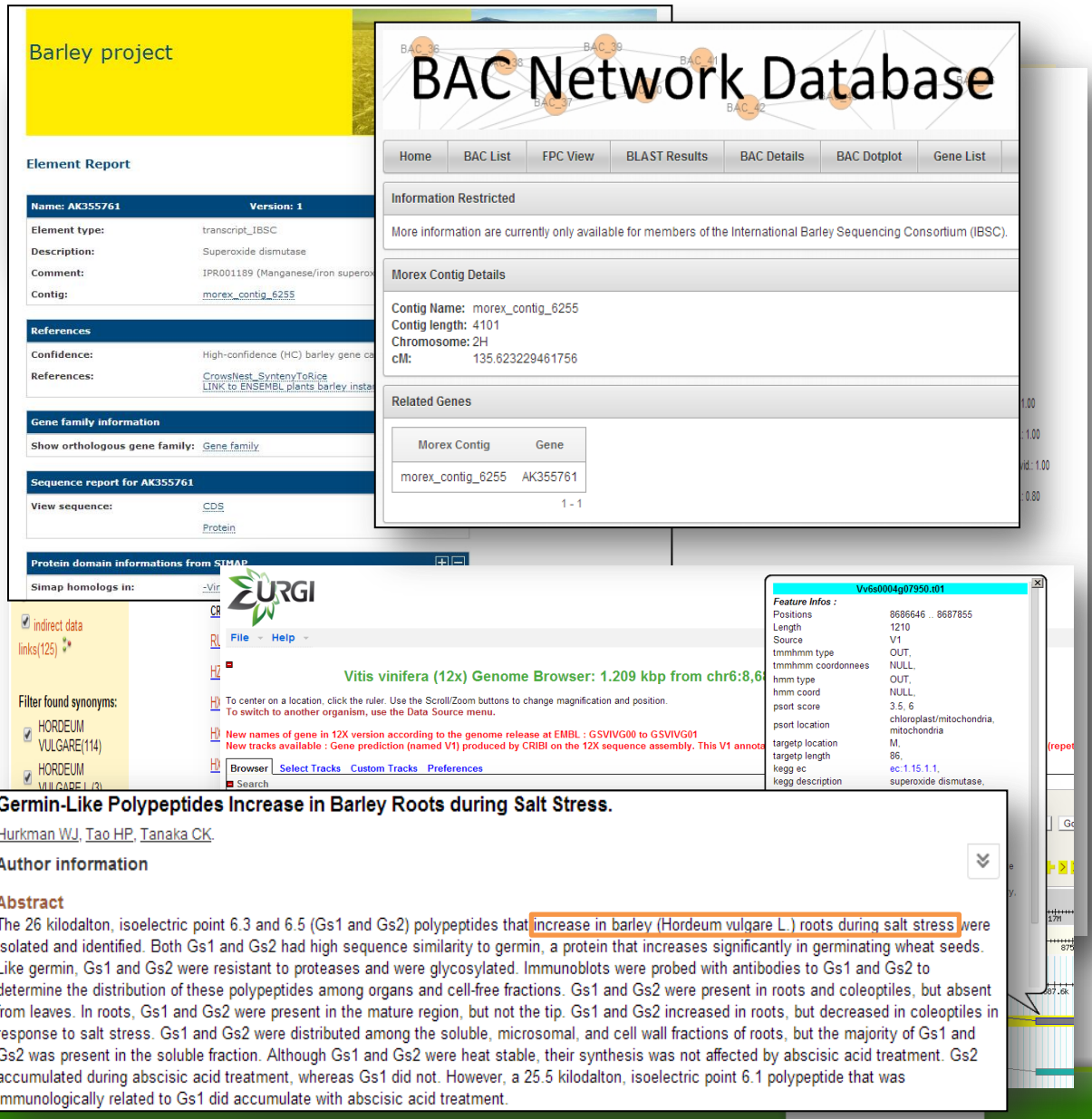
(salt OR "CG2196") AND (stress) AND (barley OR "HORDEUM VULGARE" OR …)

Get most relevant trait data:

1. D0V4H8 (Score: 51.9% - uniprot_trembl) **'response to salt stress** ...'
2. P28524 (Score: 48.9% - uniprot_sprot) '… **increase in roots during salt stress**.'

Link & rank genome annotations:

1. D0V4H8: INRA(21) EBI(3) MIPS(29) IPK(22)
2. P28524: INRA(4) MIPS(5) IPK(4) PubMed(1) BioModels(2)



Personal data and customization

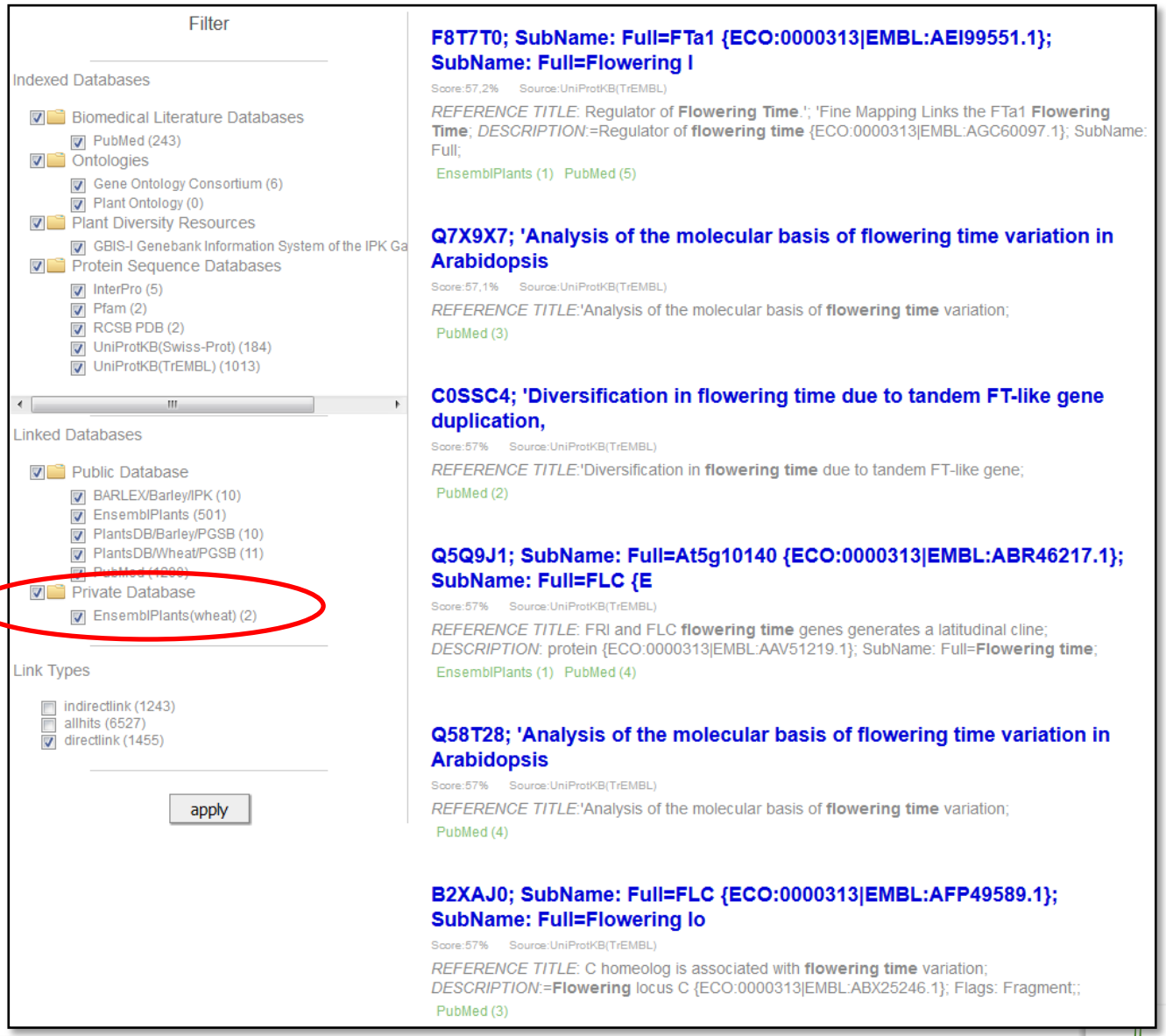
CSV format

metadata repository	metadata id	genomics datarecord ID	evidence
Pfam	PF04998	morex_contig_46112	1.0
InterPro	IPR007081	morex_contig_46112	1.0
gene_ontology	GO:0003677	morex_contig_46112	1.0
gene_ontology	GO:0003899	morex_contig_46112	1.0
gene_ontology	GO:0006351	morex_contig_46112	1.0

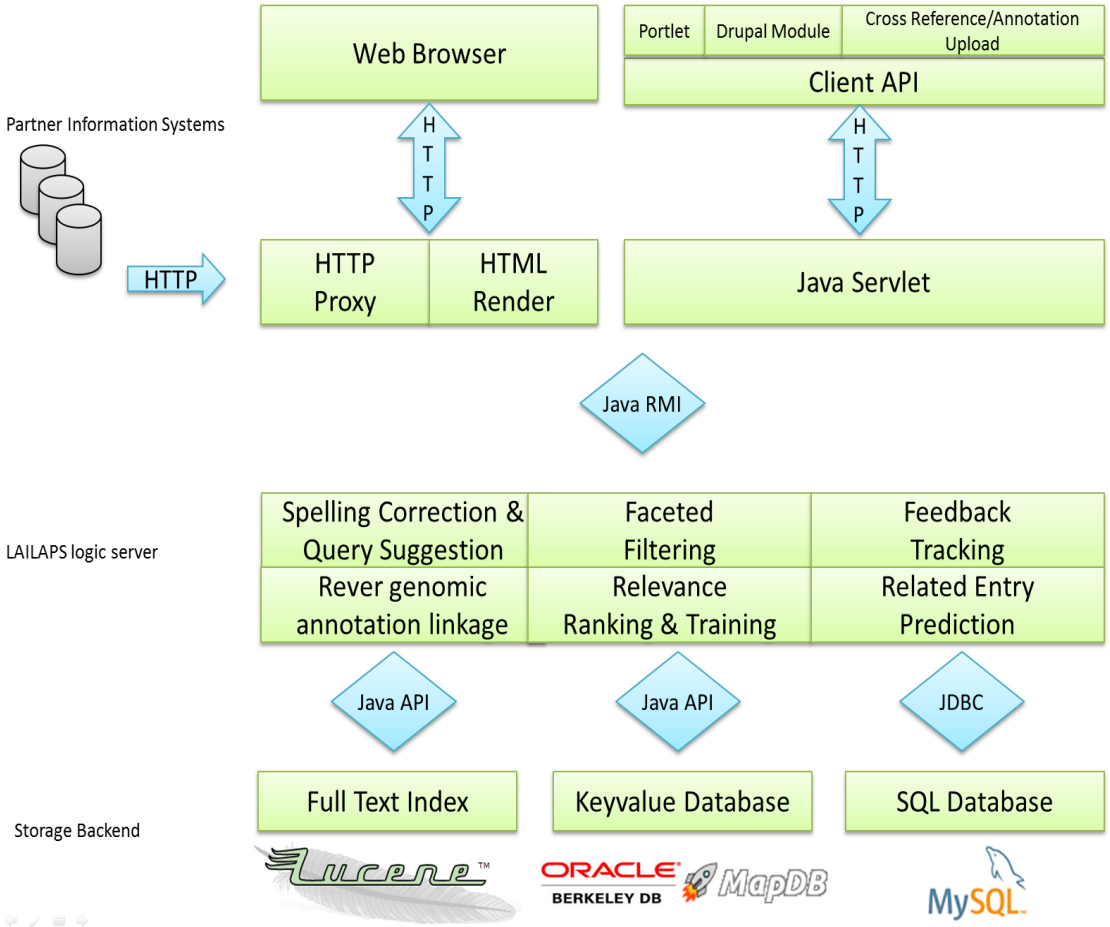
GFF3 format

1	2	3	4	5	6	7	8	9
Arad1A.1	EMBL	gene	75	1268	.	-	0	note=conserved hypothetical protein;Dbxref=UniProtKB/TrEMBL:A0A060T284;...
Arad1A.1	EMBL	exon	75	1268	.	-	0	Parent=HG937691.3
Arad1A.1	EMBL	gene	1779	2441	.	-	0	note=no similarity;Dbxref=UniProtKB/TrEMBL:A0A060SWA0;...
Arad1A.1	EMBL	exon	1779	2441	.	-	0	Parent=HG937691.5
Arad1A.1	EMBL	gene	3702	4043	.	+	0	note=similar to uniprot|P39534 Saccharomyces cerevisiae SACE0J00586g MBB1_YEAST...

Personal data and customization



Implementation



Efficient Algorithms and Data Structures:

- In-memory data structures
  - compressed HashMaps
  - bit-efficient, dynamic number types
- In-memory pre-filter
  - Bloom filter
  - Tries
- Hybrid of relational and key-value databases
- Off-heap memory (avoid garbage collection)

Hardware Requirements:

- RAM: 16GB
- CPU: 4 core, 4GHz
- storage 800GB SSD
- cost < $1500
- performance:
  - max. response time: 20 sec. (broad query: e.g. „gene“ etc.)
  - 25 parallel queries

Acknowledgements

IPK GATERSLEBEN

Member of the Leibniz Association

Matthias Lange
Daniel Arend
Christian Colmsee
Uwe Scholz
Thomas Münch
Jens Bauernfeind

LAILAPS THE PLANT SCIENCE SEARCH ENGINE

lailaps.ipk-gatersleben.de

transPLANT → from March 2015 de.NBI

Thank you for your attention

and see you at the Poster