The BEL Information Extraction Workflow (BELIEF): Updates and Evaluation

Sam Ansari
Philip Morris International R&D, Philip Morris Products S.A.
(part of Philip Morris International group of companies)

13 April 2016
Biocuration 2016 – Workshop 9: Text Mining and Biocuration: Moving to Integration

Philip Morris International is the sole source of funding and sponsor of this project.

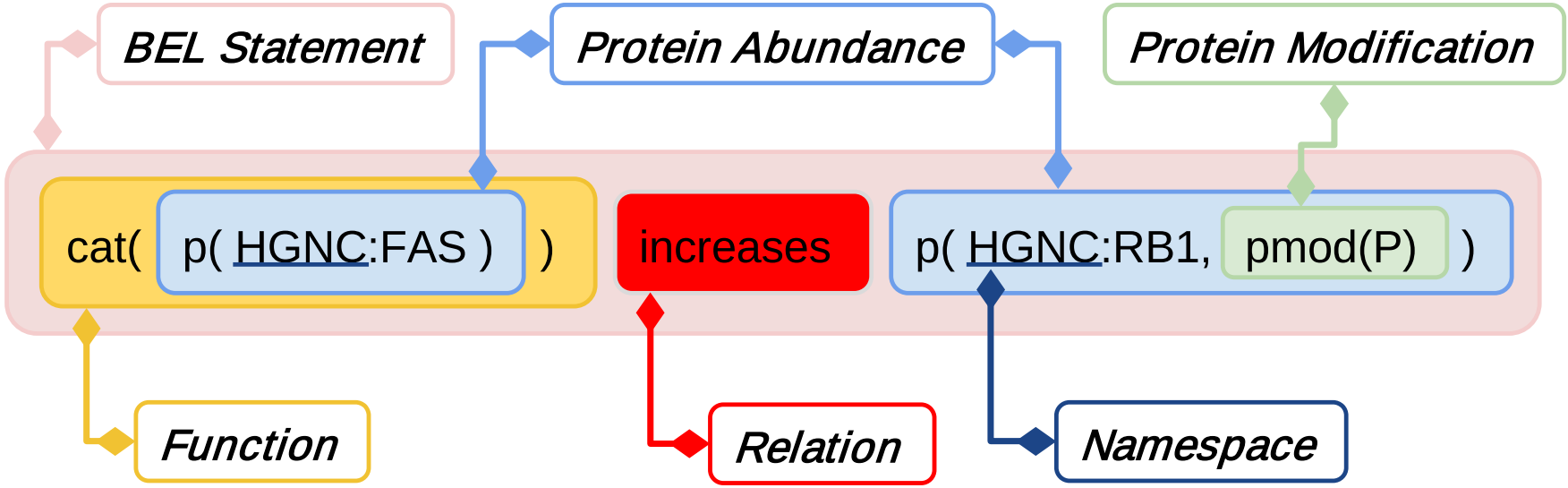
Overview

- Biological Expression Language (BEL)
- The Text Mining Pipeline (BELIEF Pipeline)
- Performance and Limitations of Text Mining
- The Curation Interface (BELIEF Dashboard)
- Performance and Limitations of the Curation Interface
- Summary

Biological Expression Language (BEL)
BEL Nanopub

Openbel.org

BEL Statement



Citation

```
SET Citation = {“PubMed”, “Regulation of Rb and E2F by signal transduction cascades:
divergent effects of JNK1 and p38 kinases.” , “EMBO J. 1999 Mar 15;18(6):1559-70.“ ,
“10075927”}
```

Support

```
SET Evidence =
        “Fas stimulation of Jurkat cells is known to induce p38 kinase and we find
         a pronounced increase in Rb phosphorylation within 30 min of Fas stimulation”
```

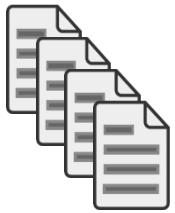
Experiment Context

```
SET Tissue = “jurkat cells”
```

Slater, T. (2014) Drug Discov. Today, 19, 193–198, Recent advances in modeling languages for pathway maps and computable biological networks.

Expectation from Text Mining
BELIEF Pipeline

Expectation



Full Text

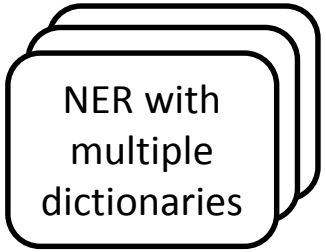
NER & Normalization

Relation Extraction

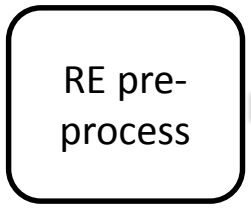
Implementation
BELIEF Pipeline

TXT

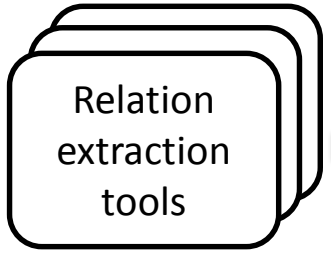
Various NLP tools



NER with multiple dictionaries



RE pre-process



Relation extraction tools

BEL Writer

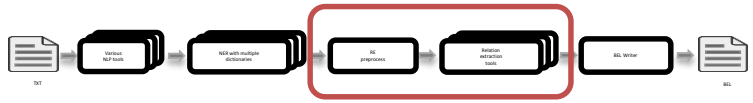
BEL

Szostak, J., Ansari, S., Madan, S., et al. (2015) Database (Oxford)., 2015, Construction of biological networks from unstructured information based on a semi-automated curation workflow.
Madan, S., Hodapp, S., Senger, P., et al. (2016) Database (Oxford)., 2016, The BEL Information Extraction Workflow (BELIEF): Evaluation in the BioCreative V BEL and IAT track (accepted)

BELIEF Pipeline – Named Entity Recognition

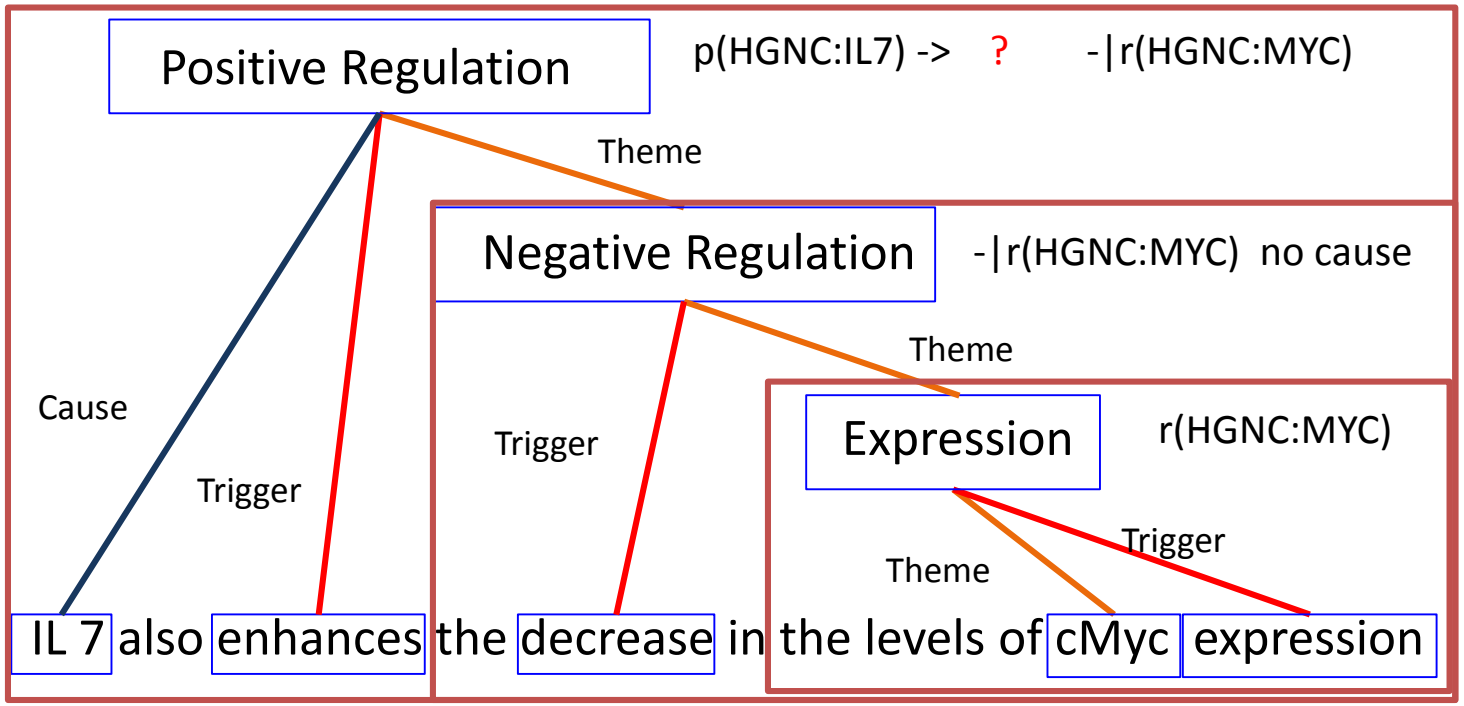
Entity class	Resources	BEL namespace
Human genes/proteins	EntrezGene/Uniprot	HGNC
Mouse genes/proteins	EntrezGene/Uniprot	MGI
Rat genes/proteins	EntrezGene/Uniprot	RGD
Protein family names	OpenBEL	SFAM
Protein complex names	OpenBEL	SCOMP
Protein complex names	Gene Ontology	GOCC
Biological processes	Gene Ontology	GOBP
Chemical names	OpenBEL	SCHEM
Chemical names	ChEBI	CHEBI
Chemical names	ChEMBL	CHEMBL
Disease names	MeSH	MESHD
Anatomical names	MeSH	MeSHAnatomy
Cell lines	Cell Line Ontology	CellLine
Cell structures	MeSH	CellStructure

Dictionary	Recall rate application adapted
Genes/Protein: (HGNC)	93 %
Chemical compounds: ChEBI	66 %
Chemical compounds: SCHEM	75 %
Chemical compounds: ChEBI + SCHEM+ ChEMBL	91 %
Selventa-human-complex	46 %
GO-Complex	64 %
Selventa-human-complex + Complex	82 %
Selventa-human-families	77 %



BELIEF Pipeline – Relation Extraction

- The BioNLP shared tasks delivers a very detailed annotation for relationship extraction similar to the information needed for BEL (TEES2.1):



- Simpler binary classification (LibLINEAR):

  IL 7 also enhances the decrease in the levels of cMyc ….

  IL7 - cMyc Relation: Yes

  p(HGNC:IL7) - - p(HGNC:MYC)

Classifies if a relation between 2 entities is existing but gives no information about the direction or type

Björne, J., Ginter, F. and Salakoski, T. (2012) BMC Bioinformatics, 13 Suppl 1, S4, University of Turku in the BioNLP'11 Shared Task

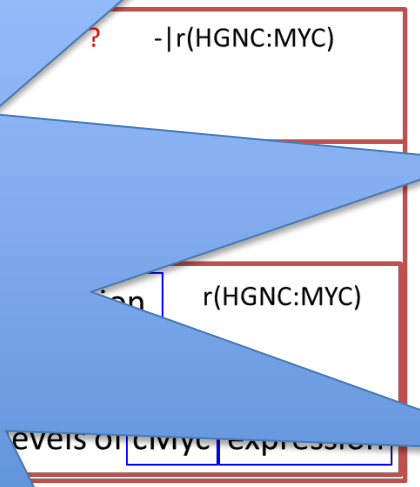
Fan, R. E., Chang, K. W. and Hsieh, C. J. (2008) J. Mach. Learn., 9, 1871–1874, LIBLINEAR: A library for large linear classification

BELIEF Pipeline - Limitations

- NLP (Sent… Detection ~6% error)
- NLP (Tokenizati… …ror)
- NER (Different Class…
- RelationExtraction (Multi…

We need manual curation!

Class	Precision	Recall	
Term	81.34	72.6…	
Function	51.16	33…	40.3…
Relationship	67.37	…1.68	43.1…
Statement	59.15	20.79	30.77

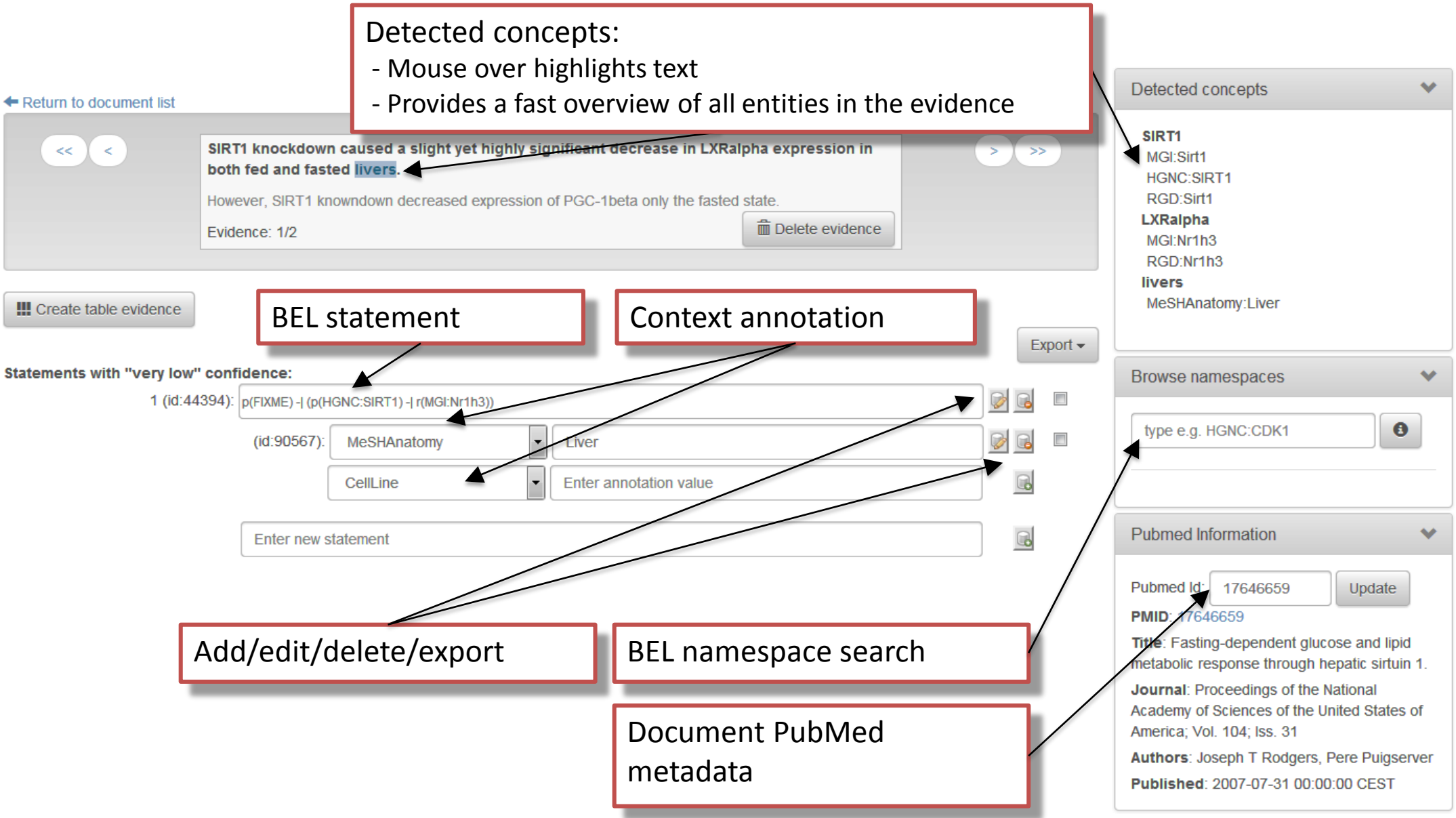


…Statem…

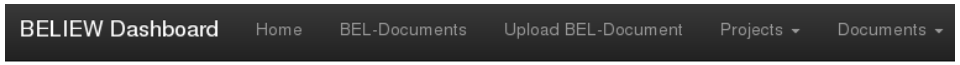
…IL7) -> p(Pla… …r) -|r(HGNC:MYC)

Ma… …tatement:
p(H… …C:IL7) -| r(HGNC:MYC)

BELIEF Dashboard – Curation Interface



BELIEF Dashboard – Other features



Create Project

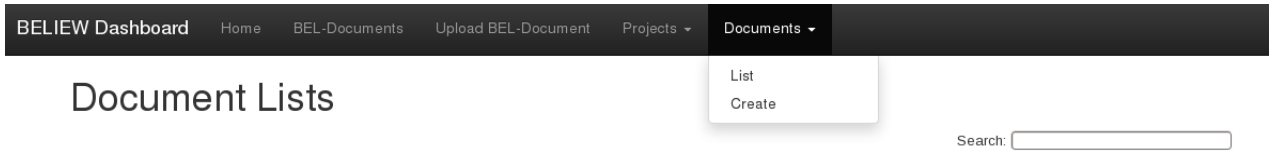
Name

Enter name

Description

Enter description

Create



Document Lists

Search:

Id	Pubmed Id
1	2234234
2	12165281
Id	Pubmed Id

Showing 1 to 2 of 2 entries

Show 10 entries

Search:

Id	Statement	EvidenceId	Evidence	Export	Curate	Delete
14580	p(HGNC:PARK7, sub(L,166,P)) -> path(MESHD:"Parkinsonian Disorders")	1	In the present study, we investigated the effects of edaravone on the neurotoxicity in PD-induced isoforms of DJ-1 containing the mutation L166P.	☑	✎	✖
14581	p(HGNC:PARK7) -- p(MESHD:"Parkinsonian Disorders")	1	In the present study, we investigated the effects of edaravone on the neurotoxicity in PD-induced isoforms of DJ-1 containing the mutation L166P.	☐	✎	✖
14582	a(CHEBI:edaravone) -> p(HGNC:SLC18A2)	2	Interestingly, our result also demonstrated that edaravone was able to up-regulate VMAT2 expression in N2a cells in a dose-dependent manner.	☑	✎	✖
14583	a(CHEBI:edaravone) -| bp(MESHPP:"Oxidative Stress")	2	Interestingly, our result also demonstrated that edaravone was able to up-regulate VMAT2 expression in N2a cells in a dose-dependent manner.	☑	✎	✖
14584	a(CHEBI:edaravone) -| path(MESHD:"Parkinsonian Disorders")	3	Our findings enhance the understanding of the neuro-protective effects of edaravone in cell models and suggest that edaravone offers significant protection in a PD-related in vitro model.	☑	✎	✖
14586	a(CHEBI:edaravone) -| bp(MESHPP:Apoptosis)	2	Interestingly, our result also demonstrated that edaravone was able to up-regulate VMAT2 expression in N2a cells in a dose-dependent manner.	☑	✎	✖
Id	Statement	EvidenceId	Evidence	Export	Curate	Delete

Showing 1 to 6 of 6 entries

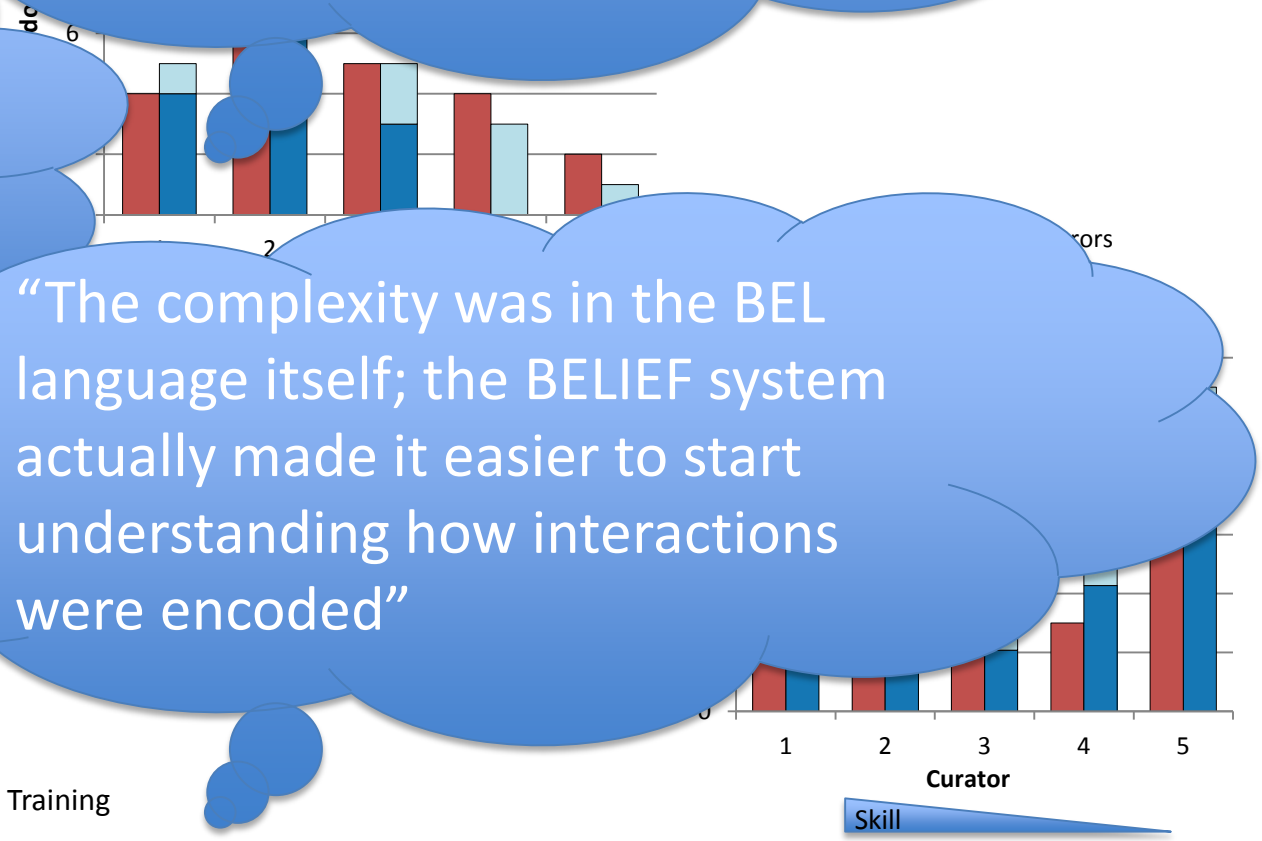
Previous 1 Next

BELIEF Dashboard – Perf
Impression

Know		
Protein		
RNA abun		
Complex abundance	3	
Biological process	10	14

"The system is very easy to learn for a user who is already familiar with BEL."

"In particular, the preselected protein identifiers were immensely useful (which I only found out when I tried to find them by hand)."



"The complexity was in the BEL language itself; the BELIEF system actually made it easier to start understanding how interactions were encoded"

Fluck, J., Madan, S., Ansari, S., et al. (2016) Submitt. to Database J. Biol. Databases curation, 2016, Training corpora for the extraction of causal relationships coded in Biological Expression Language (BEL).

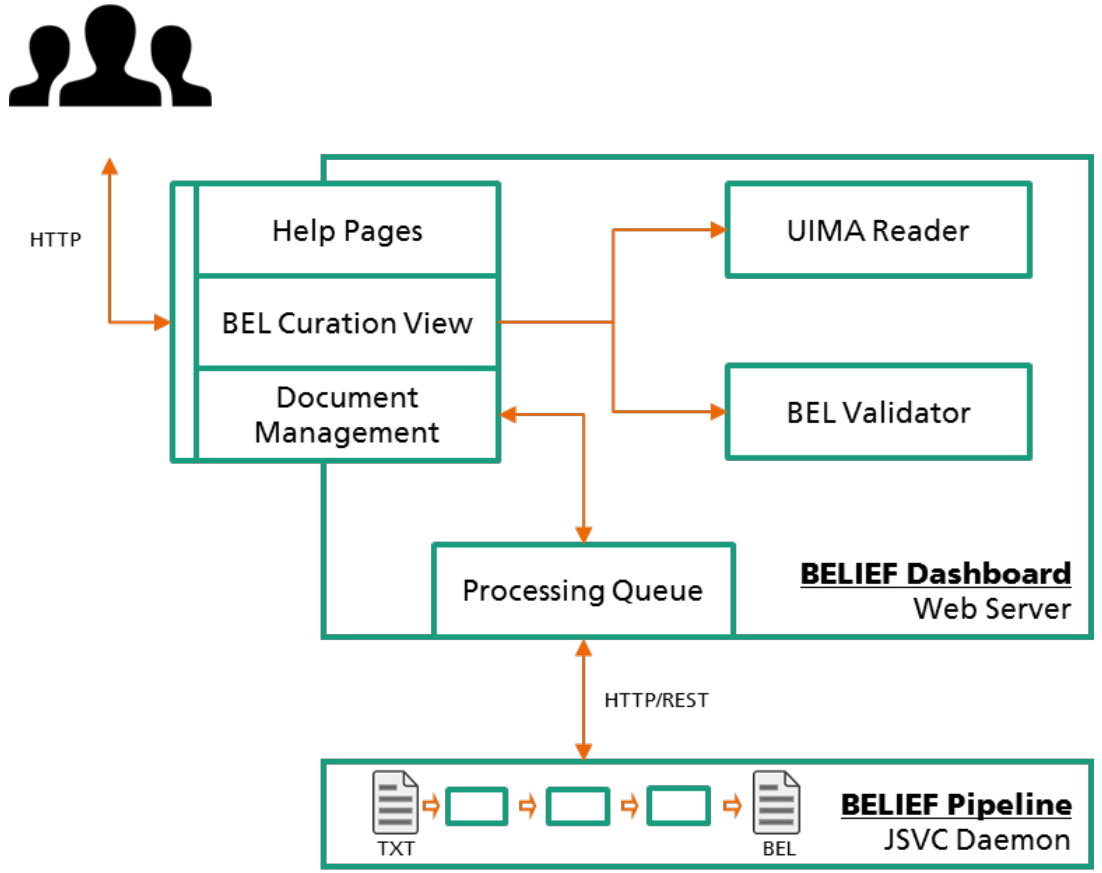
Summary

BELIEF features a text mining pipeline and a complementing curation interface with the goal to support domain experts in different stages of knowledge acquisition and network model creation

The performance of the pipeline improved and has an acceptable level

Independent tester have given BELIEF a good systems usability score and appreciate the curation interface

The overall impression of all untrained testers was that BELIEF speeds up and further simplifies the creation of BEL statements



Madan, S., Hodapp, S., Senger, P., et al. (2016) Database (Oxford)., 2016, The BEL Information Extraction Workflow (BELIEF): Evaluation in the BioCreative V BEL and IAT track (accepted)

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Summary cont'd

The new and impactful features are:

- Single point of entry including document and task management
- Reduced BEL coding effort due to full and partial BEL statement generation and validation on modifications
- Automatic citation from the Pubmed ID
- Two curation views to facilitate curation (evidence and statement centric view)
- Possibility to use custom dictionaries and re-running the text mining pipeline with these
- Show adjacent sentences to support curation

Outlook:

Additional research is foreseen to improve a series of features from workflow management to user guides.

Accessibility:

http://www.scaiview.com/belief/

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Acknowledgement

Fraunhofer SCAI

Sumit Madan
Philipp Senger
Juliane Fluck

BioCreativeV
IAT task
BEL task

William Hayes
Natalie Catlett

www.OpenBEL.org

Justyna Szostak
Marja Talikka
Julia Hoeng
Filipe Bonjour
Manuel Peitsch

PMI R&D