Exploiting text-rich descriptions for faceted discovery of web resources

María Pérez, Rafael Berlanga, Ismael Sanz, María José Aramburu
Universitat Jaume I, Castellón, Spain

Motivation

Web resource discovery through registries

Rich Metadata

Integration

Cost of publication

BioMoby

Poor Metadata

Integration

Cost of publication

BioCatalogue

Tags, categories, descriptions written in **free text**

How can this information be processed to assist in the discovery of web resources?

Task oriented web resource discovery

USER REQUIREMENTS DRIVEN

Users requirements in free text

FACETED SEARCH

Exploit the available information

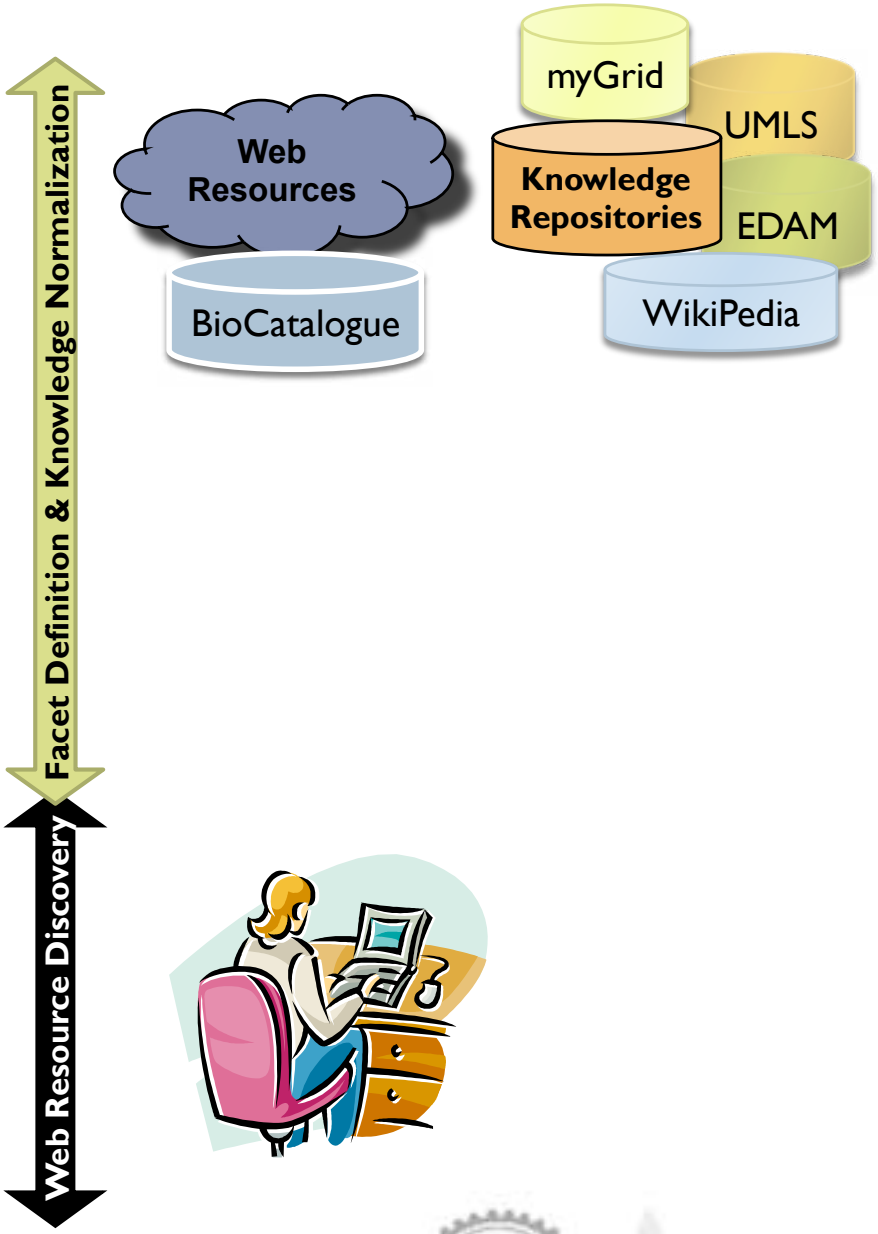
Facets defined by the users

Automatic

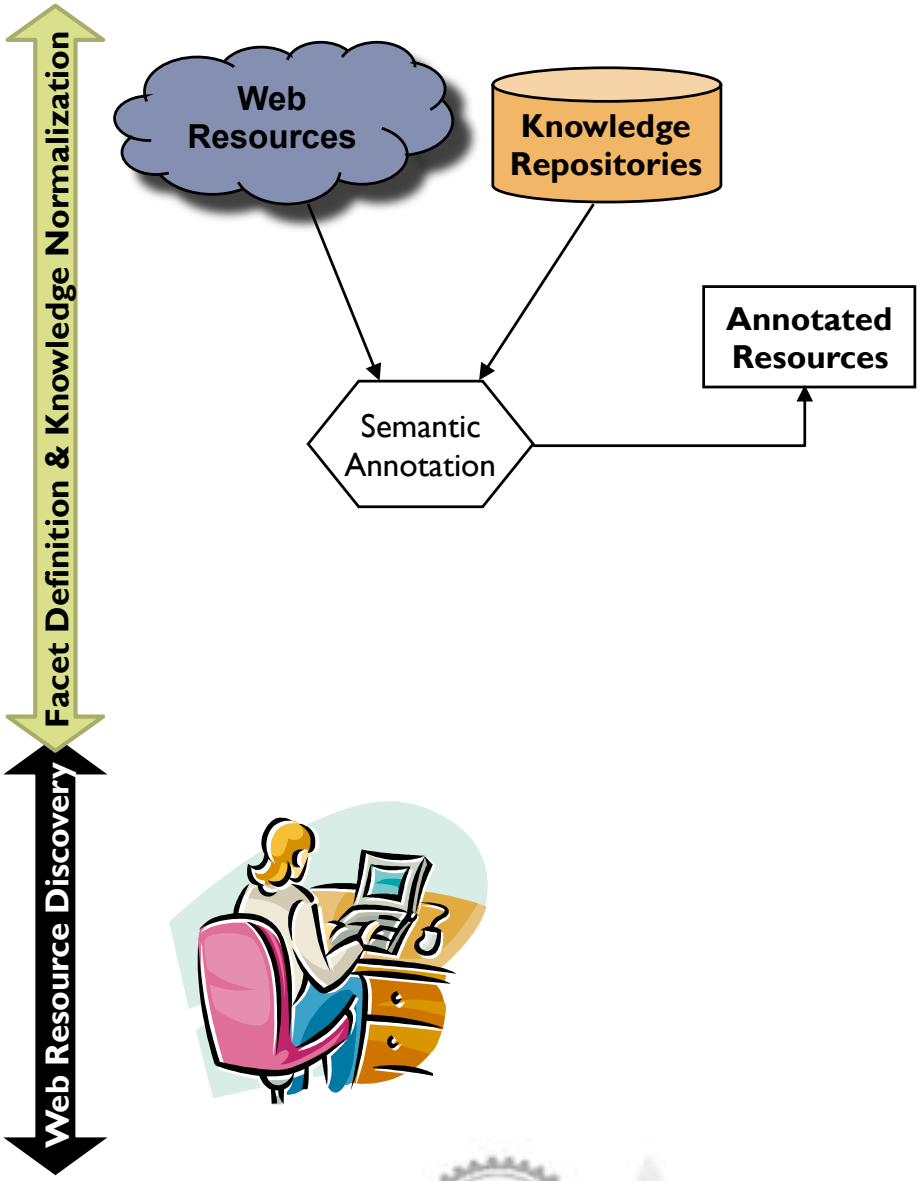
NORMALIZATION

Semantic annotation of web resources and users requirements

Automatic


Facet Definition & Knowledge Normalization
Web Resource Discovery
Web Resources
BioCatalogue
myGrid
UMLS
Knowledge Repositories
EDAM
WikiPedia

Normalization

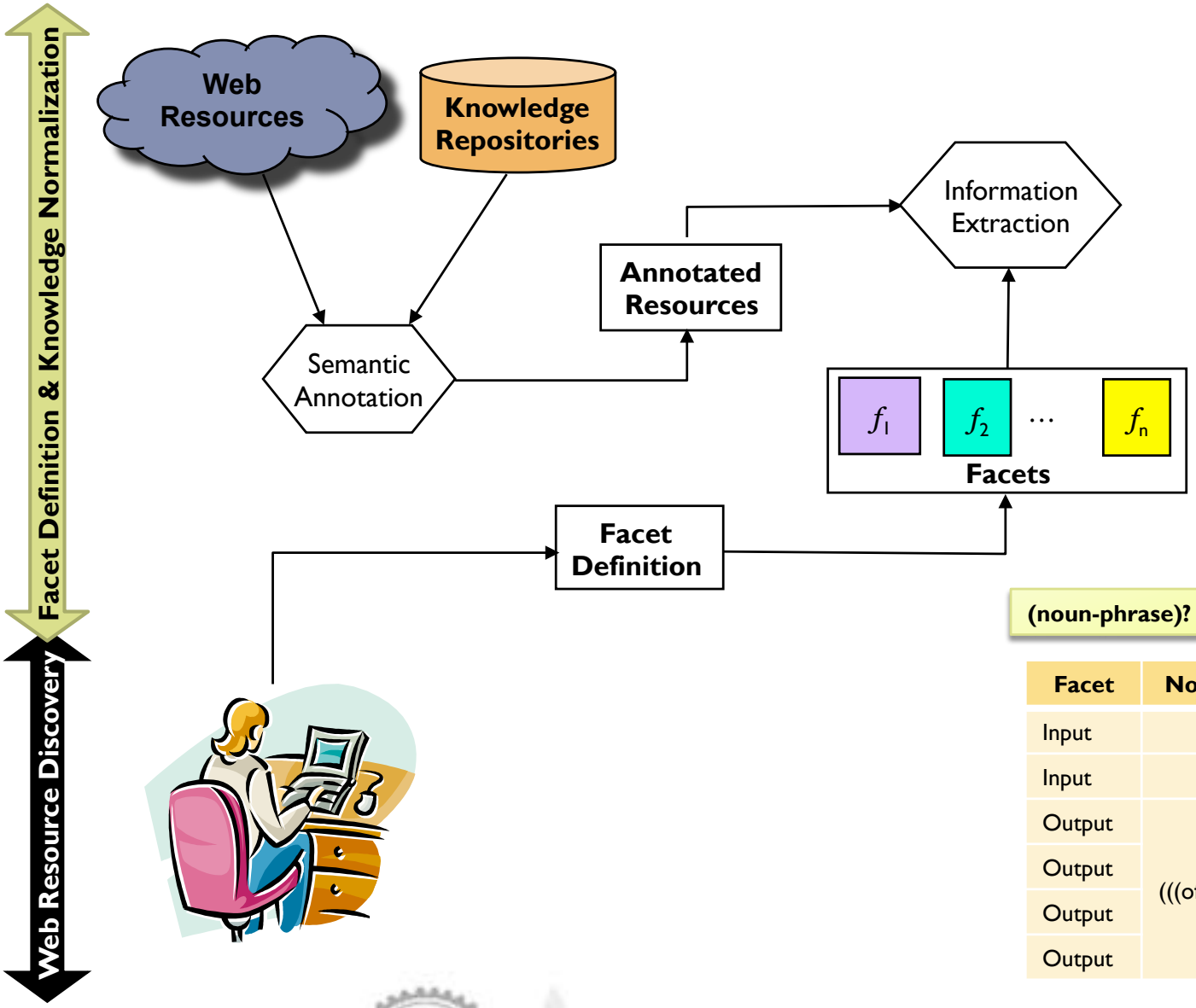


Service description

```
<Service8>
<name>Blast (DDBJ)< /name>
<category><e id="OTHR:EDAM_0001207.8 |
myGR:D9000378.15| UMLS:C0004793:T086.15| ..."> <w
id="1">Nucleotide</w> <w id="2">Sequence</w><w
id="3">Similarity</w></e></category>
<tag><e id="UMLS:C0523113.8:T059:PROC|
Wiki:W363695; 6555571;726312;4066377;4911292.10">
blast< /e>< /tag> ...
<description> <e id="UMLS:C0523113.8:T059:PROC|
Wiki:W363695; 6555571;726312;4066377;4911292.10|
OTHR:EDAM_0000646.6"> BLAST </e>
finds <e id="UMLS:C0017446.5:T083:GEOG|
UMLS:C1514562.5:T087:PRGE">regions </e> of <e
id="UMLS:C2348205:T080.8:">similarity<e> ...
< /description>...
< /service>
```

S_8={'EDAM_0001207':8, 'D9000378':15,'C0004793':15...}

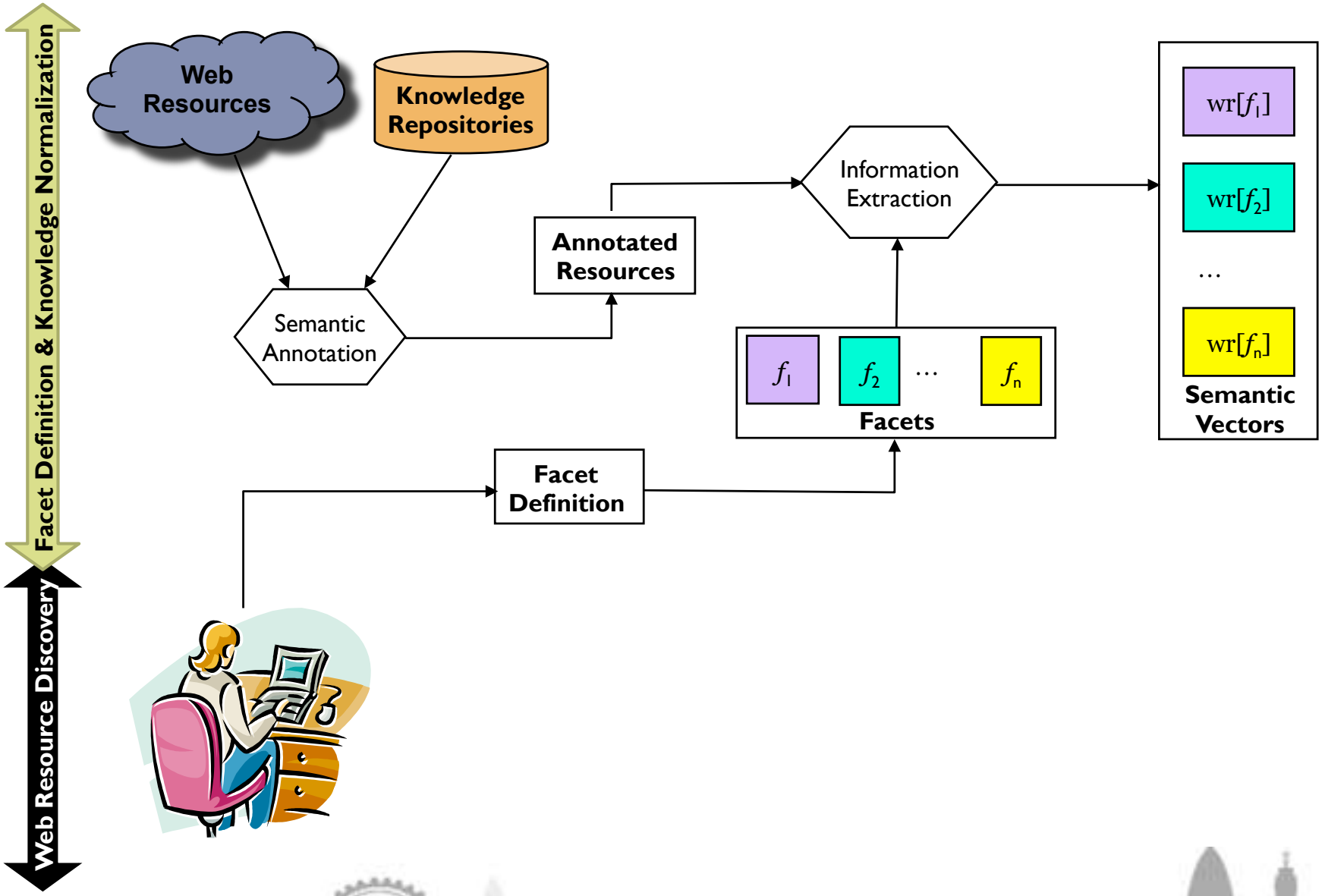
Information extraction



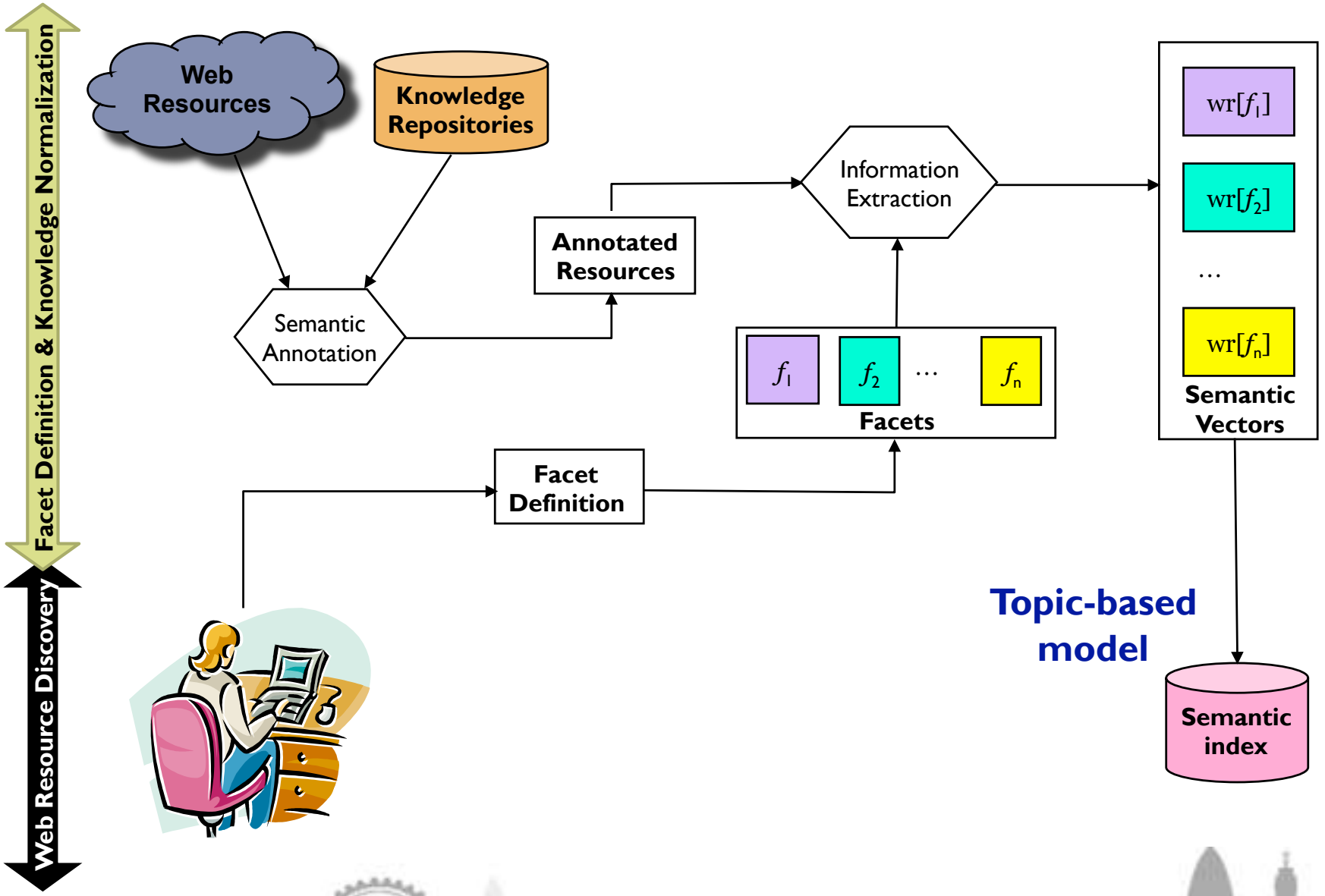
(noun-phrase)? Verb-phrase E-noun-phrase

Facet	Noun-phrase	Verb-phrase
Input	Inputs?	(is|are)
Input	-	given
Output	(((of)? E)+|it)	constructs
Output		finds
Output		retrieves
Output		returns

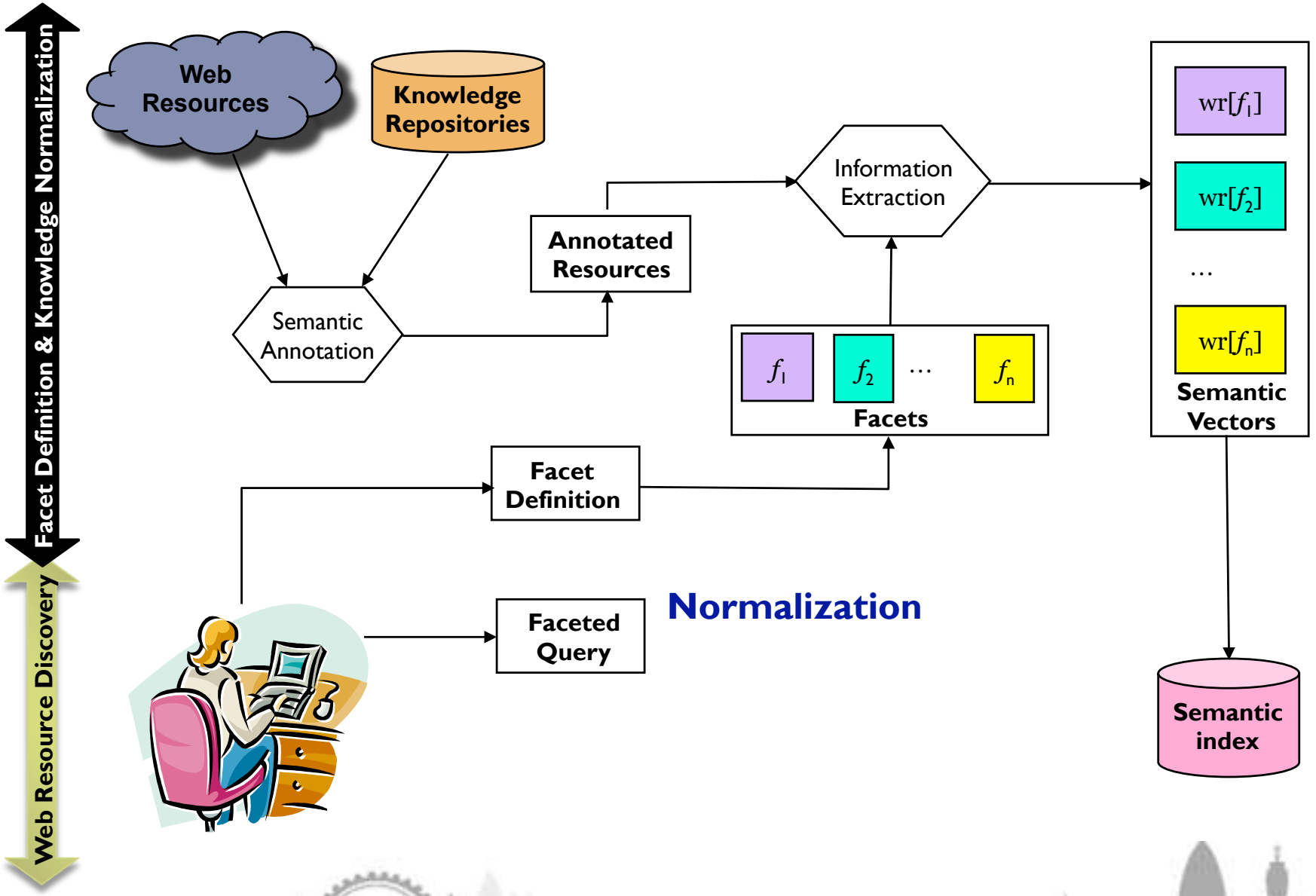
Information extraction



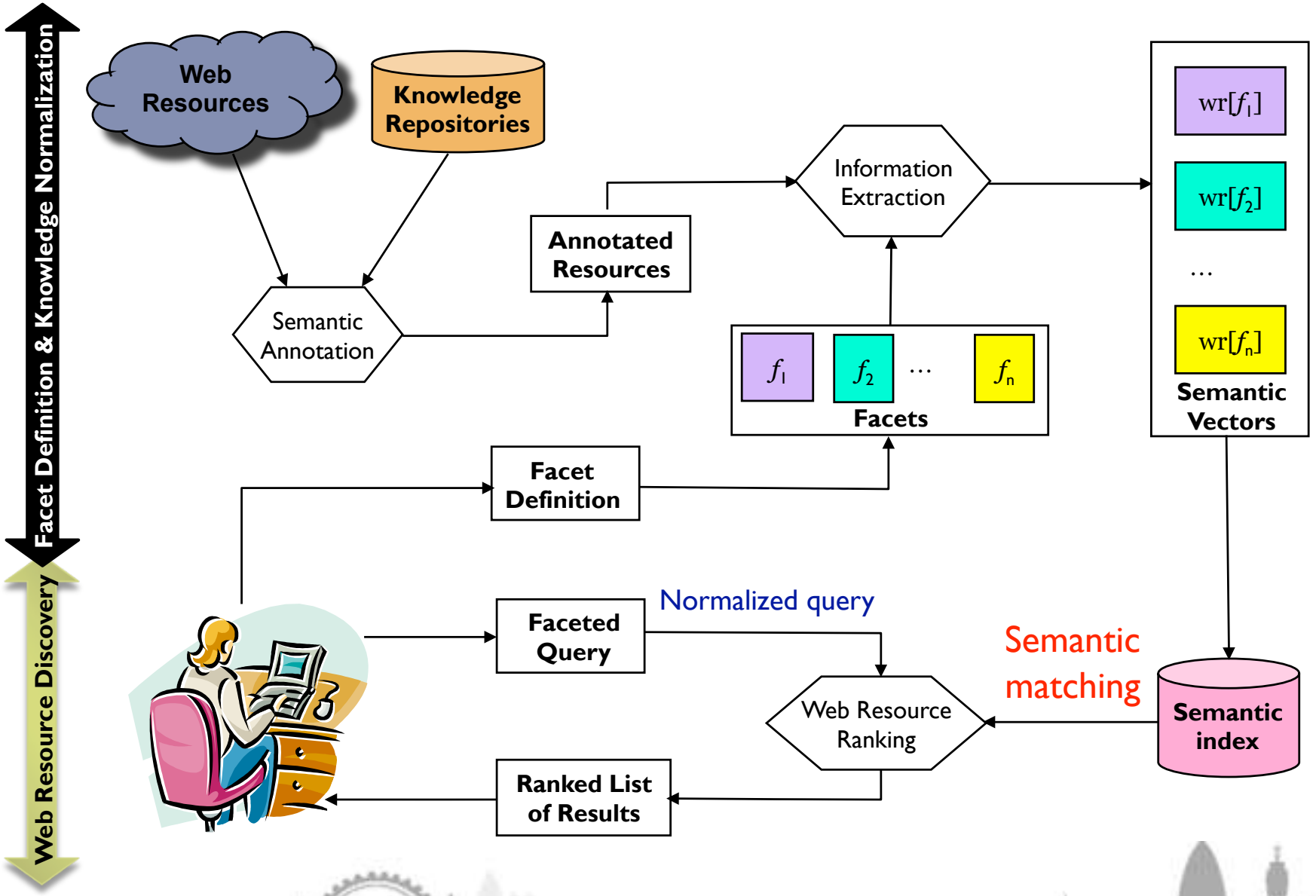
Information extraction



Web resource discovery



Web resource discovery



Web resource discovery. Results

Sequence similarity search of LRRk2 proteins

Input= ‘’ Output=‘‘

Resource	BioCatalogue categories	Input	Output	Normalized Score
MRS Blast Service	Nucleotide Sequence similarity, Protein Sequence Similarity	N/A	N/A	1
HMMER	Protein Pairwise Alignment Protein Sequence Similarity Domains, Motifs, Repeats	N/A	proteins	0,5441
blastProDom	Nucleotide Sequence similarity, Protein Sequence Similarity	protein sequence, nucleotide	blast, families	0,1407
phosphoELMdb	Protein Sequence Similarity Protein interaction Motifs, Ligand interaction	N/A	proteins	0,0584
WSScan PS	Protein Sequence Similarity	N/A	N/A	0,0576
PSI-BLAST at SIB	Protein Sequence Similarity	protein sequence	blast output	0,0433
BlastService	Protein Sequence Similarity Nucleotide Sequence Similarity	N/A	N/A	0,0421
Blast_Service	Protein Sequence Similarity Nucleotide Sequence Similarity	N/A	N/A	0,0341
tBLASTn at SIB	Protein Sequence Similarity	protein sequence	blast output	0,0313
AnalysisService	Protein Sequence Similarity	N/A	N/A	0,0274

Web resource discovery. Results

Sequence similarity search of LRRk2 proteins

Input= ‘protein sequence’ Output=‘blast'

Resource	BioCatalogue categories	Input	Output	Normalized Score
WSBlastpgp	Protein Sequence Similarity Sequence Analysis Protein Sequence Analysis	protein sequence, structure, id...	blast, format, protein sequence	1
Blast (DDBJ)	Nucleotide Sequence similarity, Protein Sequence Similarity	protein sequence, fasta, database	blast, sequence	0,0405
blastProDom	Nucleotide Sequence similarity, Protein Sequence Similarity	protein sequence, nucleotide	blast, families	0,0013
BlastbtService	Nucleotide Sequence similarity, Protein Sequence Similarity	protein sequence, database, fasta	blast, database, information, protein	0,0003
BlastDemo	Nucleotide Sequence similarity, Protein Sequence Similarity	protein sequence	blast	0,0001
TimeLogicBlastService	Nucleotide Sequence similarity, Protein Sequence Similarity	protein, blast, nucleotide	blast	$4.37 \cdot 10^{-9}$
ClustalW	Nucleotide Multiple Alignment, Protein Multiple Alignment	fasta format, sequence	sequence alignment	$2.3 \cdot 10^{-9}$
MRS Blast Service	Nucleotide Sequence similarity, Protein Sequence Similarity	N/A	N/A	$4.09 \cdot 10^{-67}$
HMMER	Protein Pairwise Alignment Protein Sequence Similarity Domains, Motifs, Repeats	N/A	proteins	$2.22 \cdot 10^{-67}$
phosphoELMdb	Protein Sequence Similarity Protein interaction Motifs, Ligand interaction	N/A	proteins	$2.37 \cdot 10^{-68}$

Conclusions

A faceted discovery of web resources

SEMANTIC ANNOTATION

FACETED SEARCH

- Exploits the available information
- Personalizable facets and topics

→ Self-configurable system

Questions?