Semi-automatic Population of a Pharmaceutical Company Ontology

Master Thesis

Ilinca Tudose

November 2, 2012

Supervisor Prof. Dr. Hannah Bast

Outline

1. Motivation
2. Contribution
3. System Overview
4. Triplet extraction
5. Conflicts
6. Evaluation
7. Patent Use Case
8. Further Work
9. Conclusion

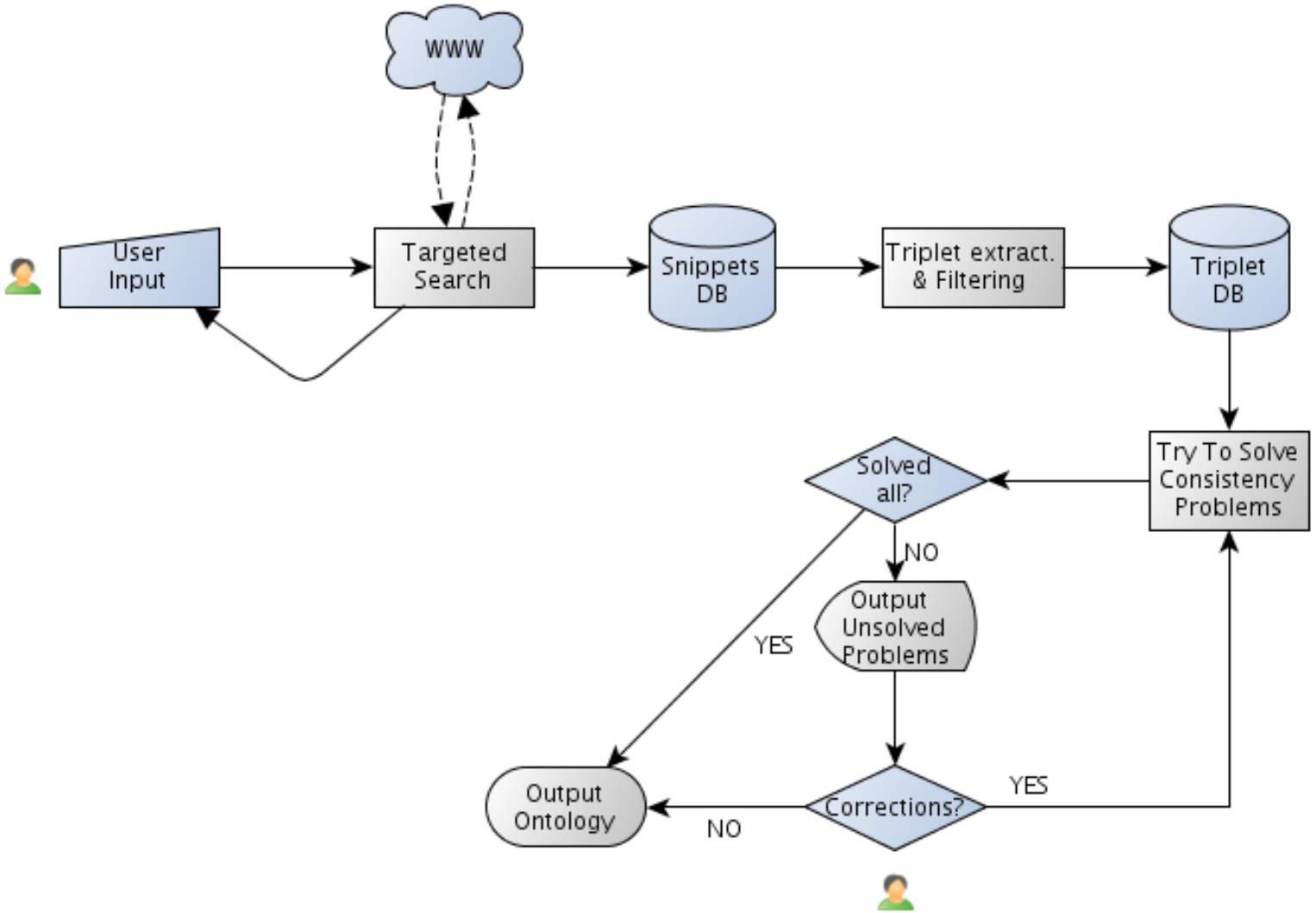
Motivation

- Competitive intelligence
  - Take informed decisions and to correctly evaluate risks, markets, opportunities, their and their competitors weaknesses and strengths
  - *Who* is my competitor?
  - The lack of a M&A ontology
- Ontology structure
- 'Googling' approach
  - 90% of the information needed for competitive intelligence could be found on the Internet (Teo and Choo, 2001)
  - “collective knowledge is much more powerful than individual knowledge” (Cimiano and Staab, 2004)

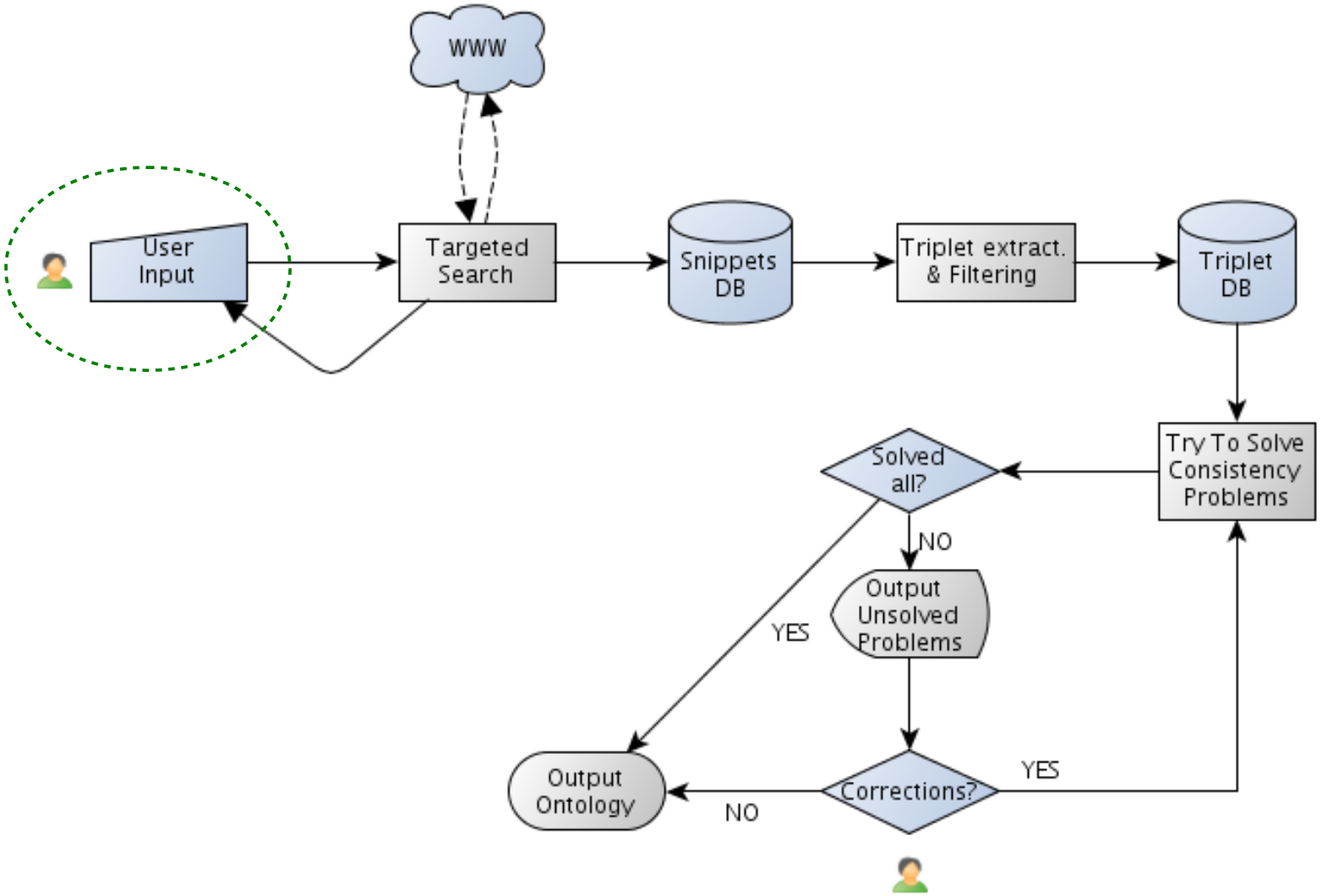
Contribution

- Information acquisition from the WWW
- Company recognition (training)
- Pattern based relation recognition
- Company synonyms and preferred terms
- Filtering mistakes
- Consistent ontology building
- Evaluation

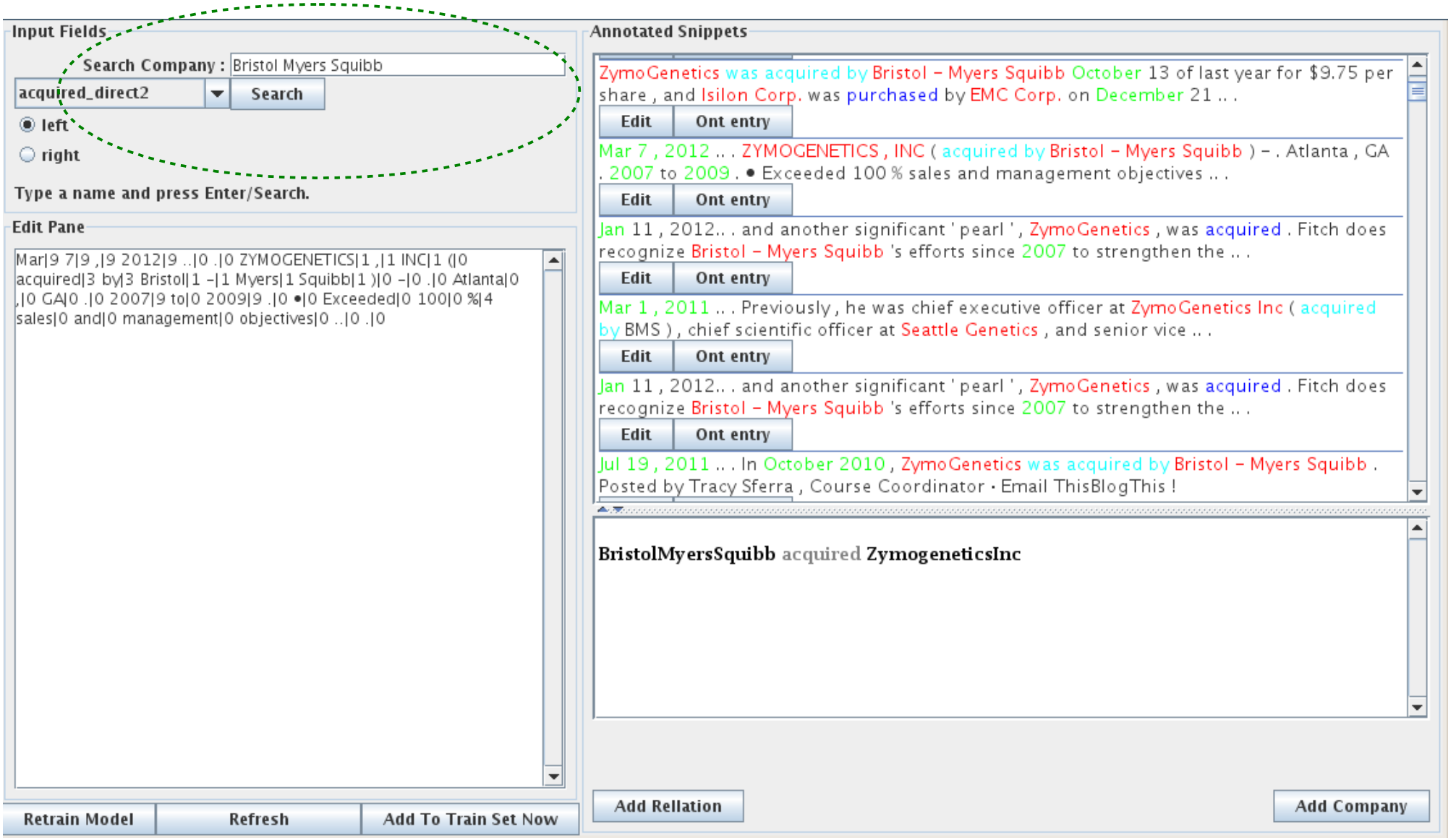
Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

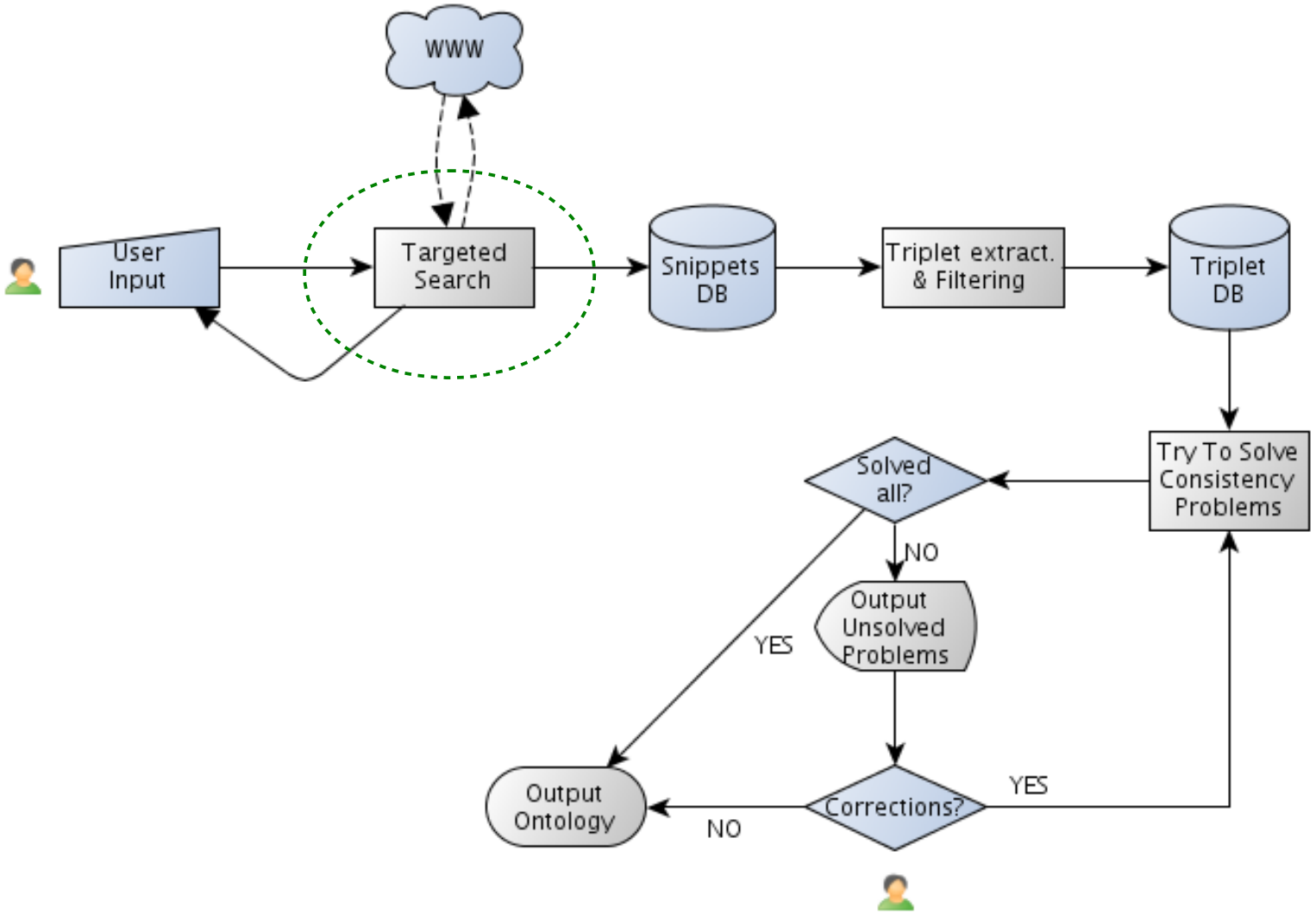
Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

Overview

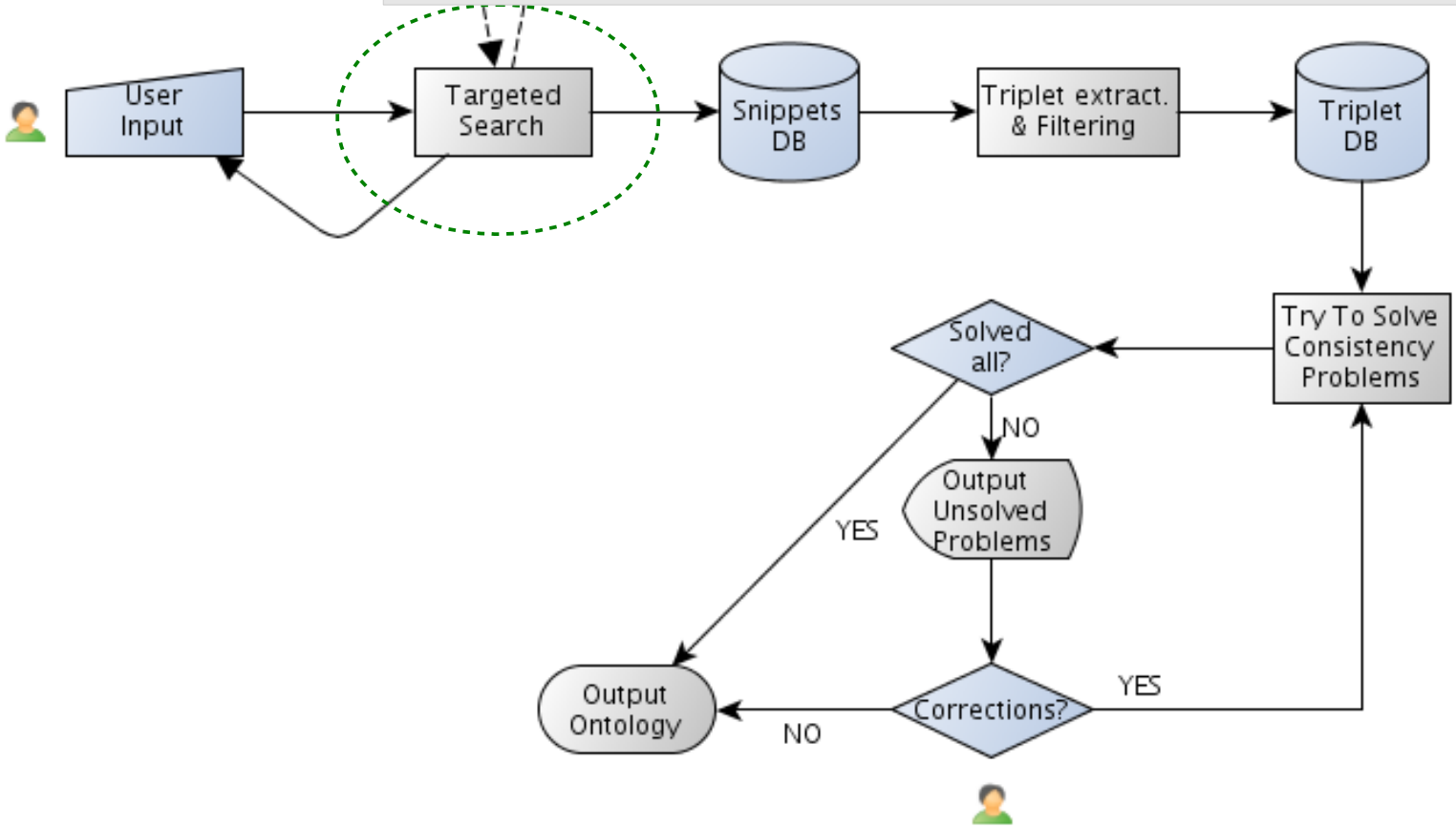

Input Fields
Search Company : Bristol Myers Squibb
acquired_direct2
Search
left
right
Type a name and press Enter/Search.
Edit Pane
Mar|9 7|9 ,|9 2012|9 ..|0 .|0 ZYMOGENETICS|1 ,|1 INC|1 (|0 acquired|3 by|3 Bristol|1 -|1 Myers|1 Squibb|1)|0 -|0 .|0 Atlanta|0 ,|0 GA|0 .|0 2007|9 to|0 2009|9 .|0 •|0 Exceeded|0 100|0 %|4 sales|0 and|0 management|0 objectives|0 ..|0 .|0
Retrain Model
Refresh
Add To Train Set Now
Annotated Snippets
ZymoGenetics was acquired by Bristol - Myers Squibb October 13 of last year for $9.75 per share , and Isilon Corp. was purchased by EMC Corp. on December 21 .. .
Edit
Ont entry
Mar 7 , 2012 .. . ZYMOGENETICS , INC (acquired by Bristol - Myers Squibb) - . Atlanta , GA . 2007 to 2009 . • Exceeded 100 % sales and management objectives .. .
Edit
Ont entry
Jan 11 , 2012.. . and another significant ' pearl ' , ZymoGenetics , was acquired . Fitch does recognize Bristol - Myers Squibb 's efforts since 2007 to strengthen the .. .
Edit
Ont entry
Mar 1 , 2011 .. . Previously , he was chief executive officer at ZymoGenetics Inc (acquired by BMS) , chief scientific officer at Seattle Genetics , and senior vice .. .
Edit
Ont entry
Jan 11 , 2012.. . and another significant ' pearl ' , ZymoGenetics , was acquired . Fitch does recognize Bristol - Myers Squibb 's efforts since 2007 to strengthen the .. .
Edit
Ont entry
Jul 19 , 2011 .. . In October 2010 , ZymoGenetics was acquired by Bristol - Myers Squibb . Posted by Tracy Sferra , Course Coordinator · Email ThisBlogThis !
BristolMyersSquibb acquired ZymogeneticsInc
Add Relation
Add Company

Overview

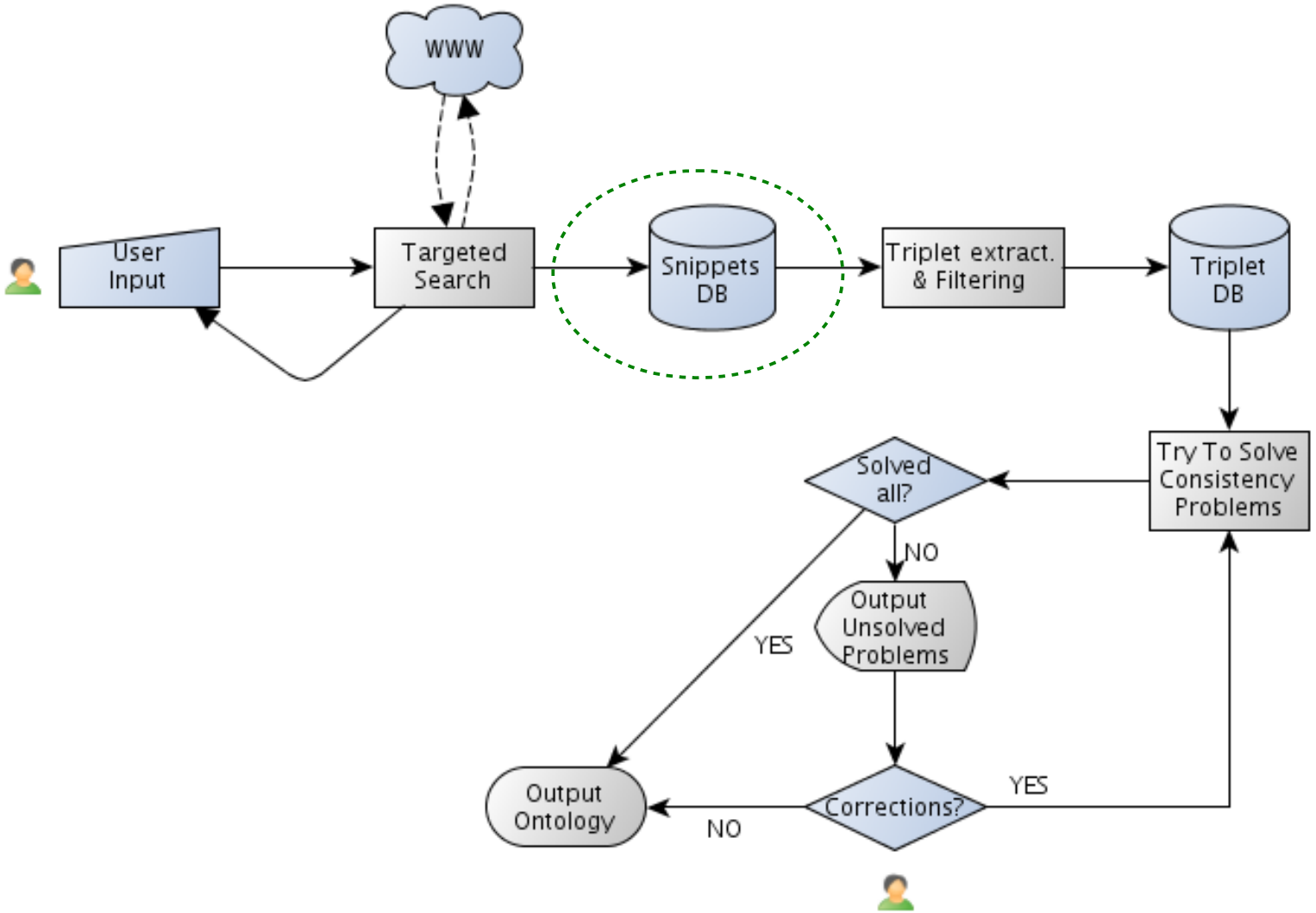

WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

Overview

Bristol Myers Squibb signed definitive agreement to acquire
Bristol Myers Squibb acquired
Bristol Myers Squibb completed the acquisition



Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

Overview

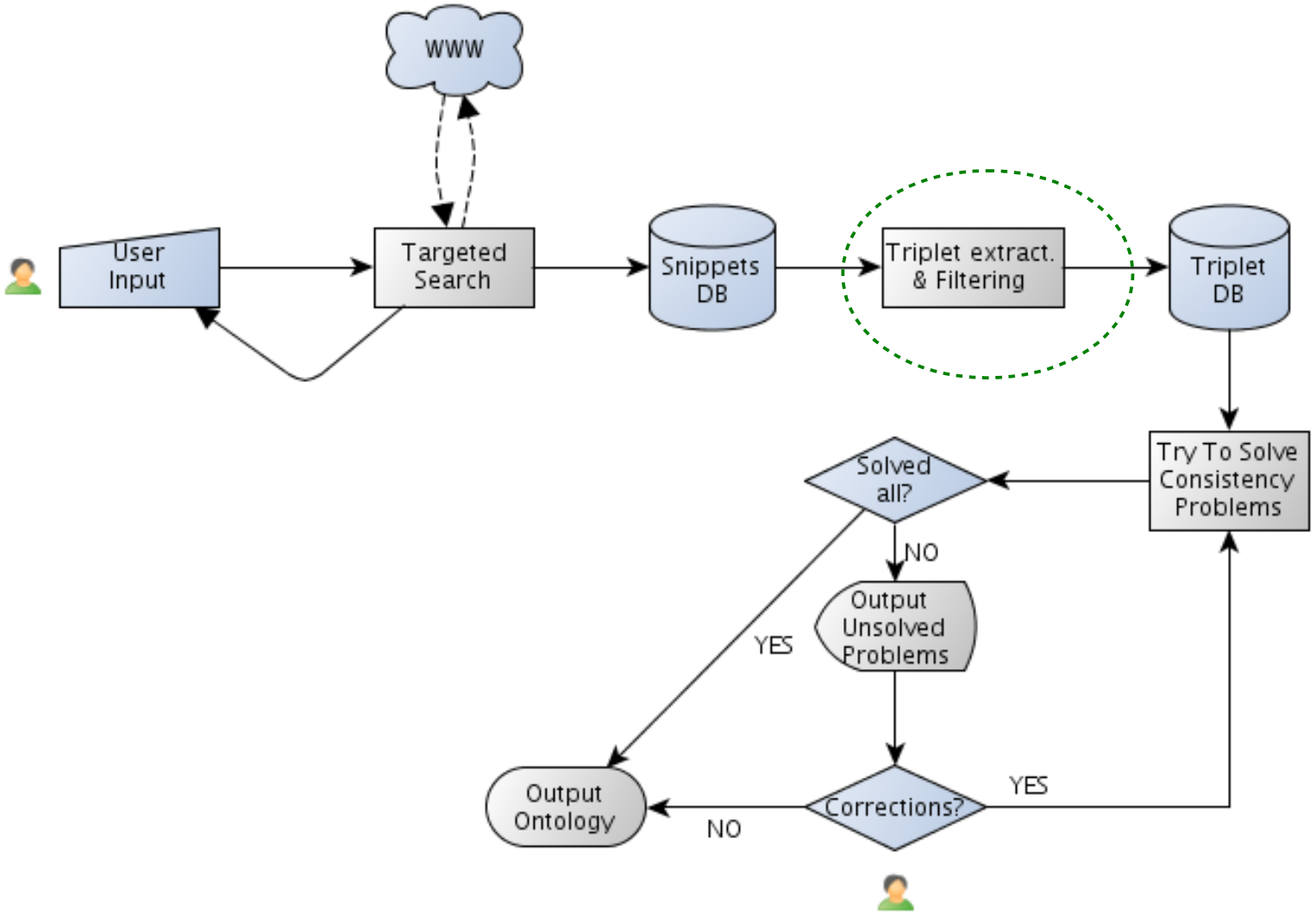
Novartis AG - Knowmore.org
knowmore.org/wiki/index.php?title=Novartis_AG
8 Jul 2011 – The Company is based in Basel, Switzerland. In July 2008, **Novartis AG acquired** a 25% stake in Alcon, Inc. from Nestle SA.

Novartis AG | Private Company Financial Research | PrivCo.com
www.privco.com/private-company/novartis-ag
PrivCo's M&A Activity table for Novartis AG displays the mergers and acquisitions involving Novartis AG, for example if **Novartis AG acquired** or was acquired by **...**

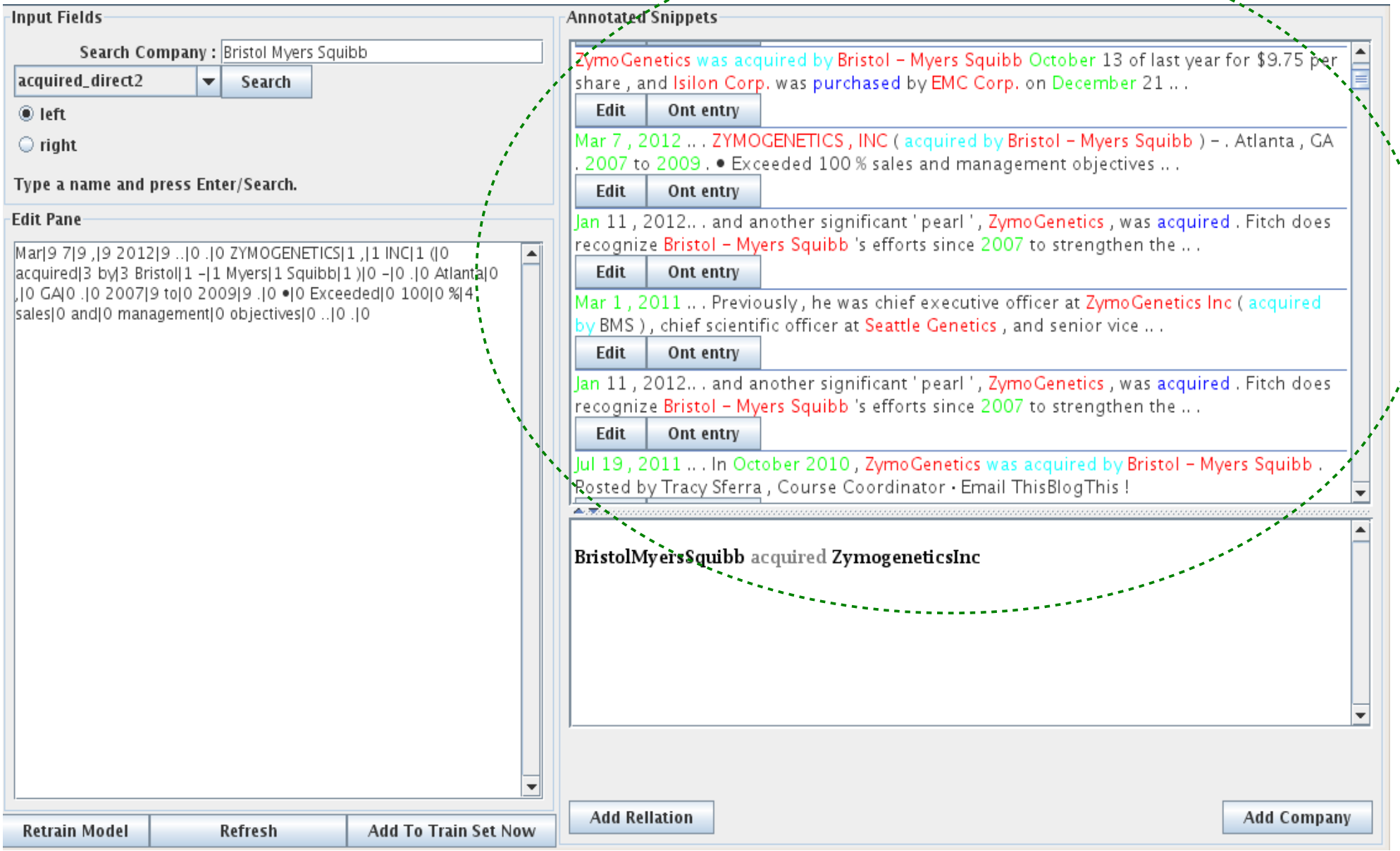
Alcon - Advanced Vision Technologies of Golden, Colorado
www.avtlens.com/#!alcon
Novartis AG acquired its 77 percent majority ownership in a two step transaction with Nestlé S.A. In July 2008, Novartis purchased from Nestlé S.A. **...**

"Switzerland's Novartis AG" Search - The Business Journals
www.bizjournals.com/search?q=%22Switzerland's+Novartis...
When Switzerland's **Novartis AG acquired** Chiron Corp. in 2005, investors, city planners, and Chiron's then-2,300 local employees were wondering what would **...**

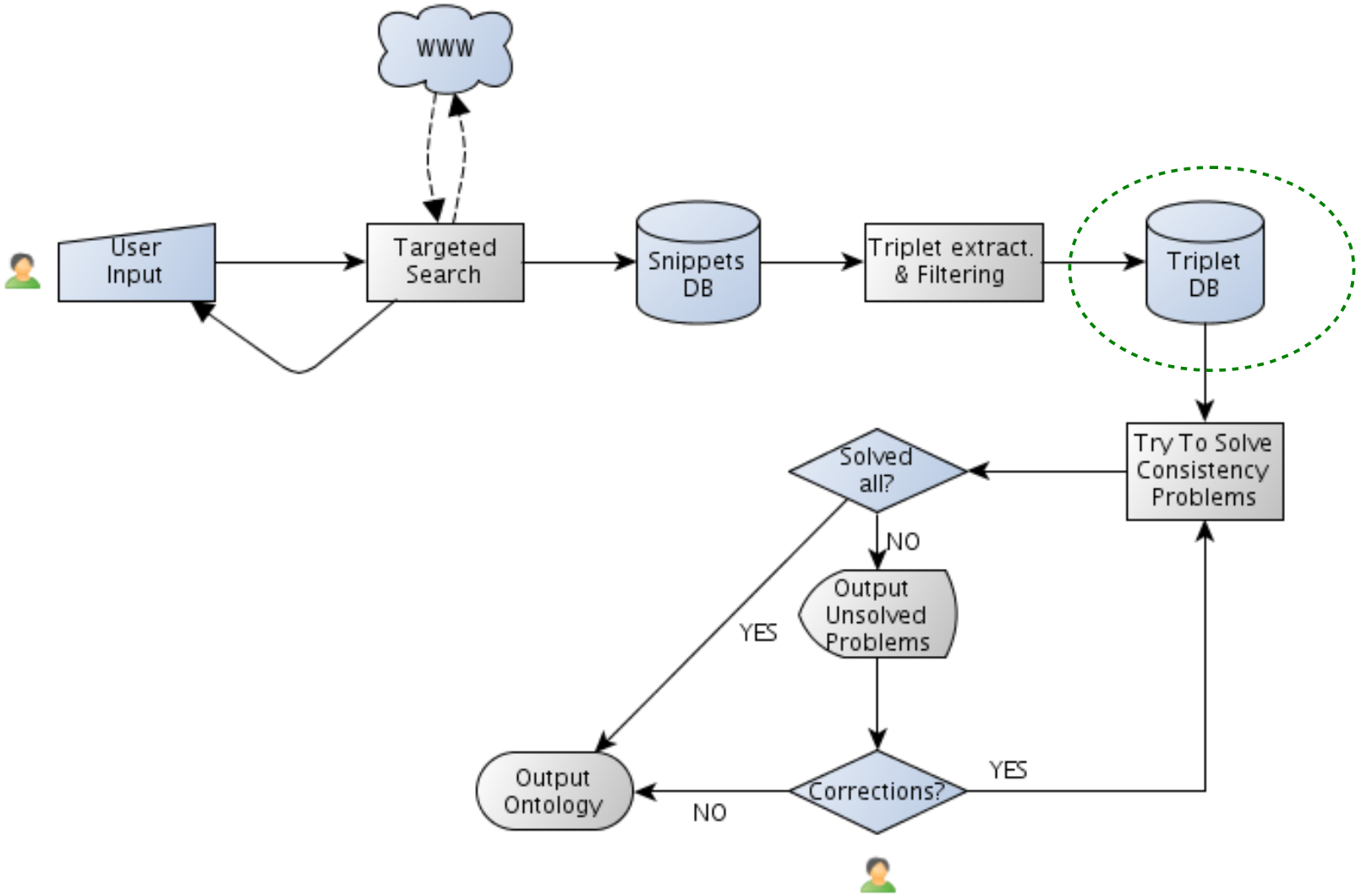
Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

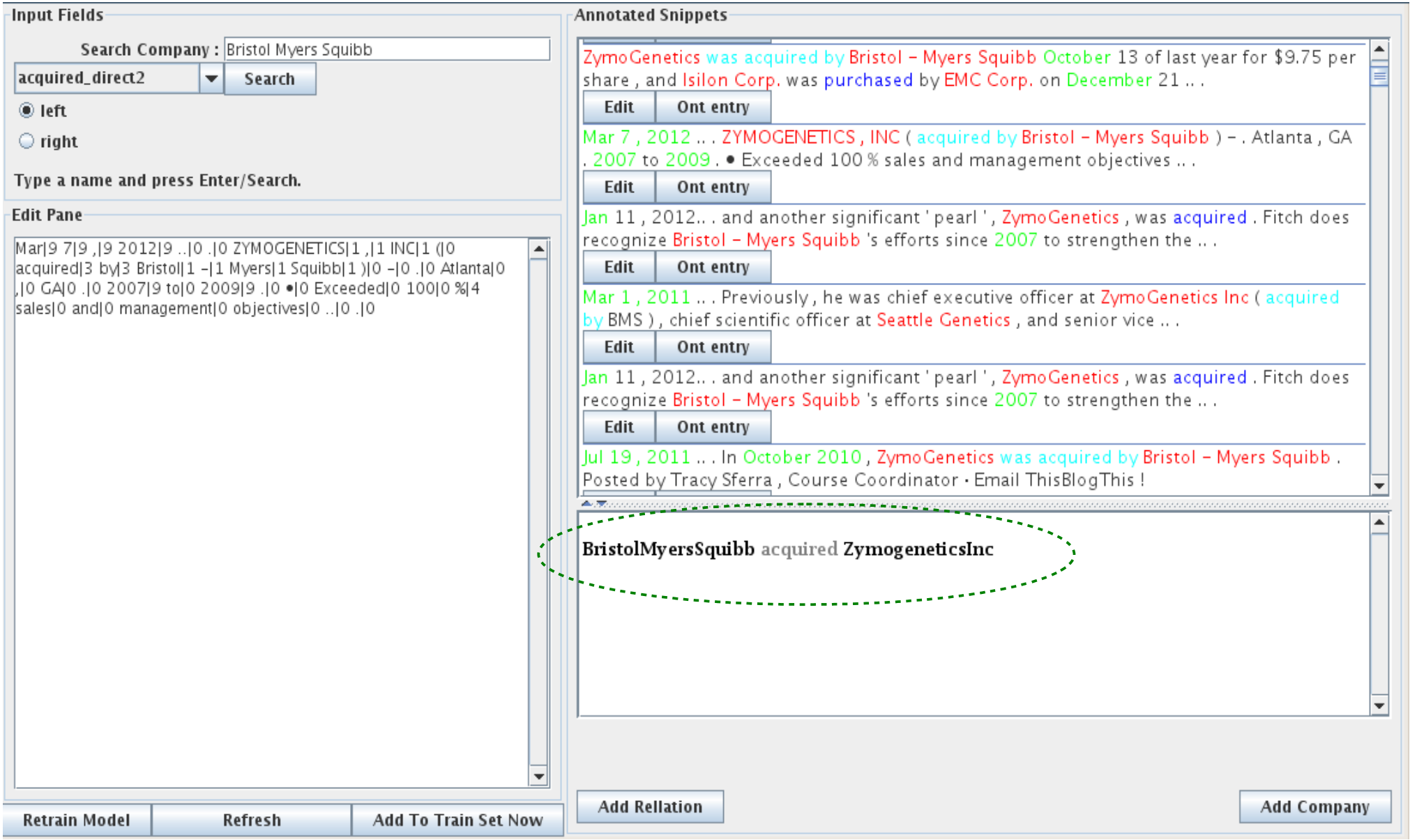
Overview


Input Fields
Search Company : Bristol Myers Squibb
acquired_direct2
Search
left
right
Type a name and press Enter/Search.
Edit Pane
Mar|9 7|9 ,|9 2012|9 ..|0 .|0 ZYMOGENETICS|1 ,|1 INC|1 (|0 acquired|3 by|3 Bristol|1 -|1 Myers|1 Squibb|1)|0 -|0 .|0 Atlanta|0 ,|0 GA|0 .|0 2007|9 to|0 2009|9 .|0 •|0 Exceeded|0 100|0 %|4 sales|0 and|0 management|0 objectives|0 ..|0 .|0
Retrain Model
Refresh
Add To Train Set Now
Annotated Snippets
ZymoGenetics was acquired by Bristol - Myers Squibb October 13 of last year for $9.75 per share , and Isilon Corp. was purchased by EMC Corp. on December 21 .. .
Edit
Ont entry
Mar 7 , 2012 .. . ZYMOGENETICS , INC (acquired by Bristol - Myers Squibb) - . Atlanta , GA . 2007 to 2009 . • Exceeded 100 % sales and management objectives .. .
Edit
Ont entry
Jan 11 , 2012.. . and another significant ' pearl ' , ZymoGenetics , was acquired . Fitch does recognize Bristol - Myers Squibb 's efforts since 2007 to strengthen the .. .
Edit
Ont entry
Mar 1 , 2011 .. . Previously , he was chief executive officer at ZymoGenetics Inc (acquired by BMS) , chief scientific officer at Seattle Genetics , and senior vice .. .
Edit
Ont entry
Jan 11 , 2012.. . and another significant ' pearl ' , ZymoGenetics , was acquired . Fitch does recognize Bristol - Myers Squibb 's efforts since 2007 to strengthen the .. .
Edit
Ont entry
Jul 19 , 2011 .. . In October 2010 , ZymoGenetics was acquired by Bristol - Myers Squibb . Rosted by Tracy Sferra , Course Coordinator · Email ThisBlogThis !
BristolMyersSquibb acquired ZymogeneticsInc
Add Rellation
Add Company

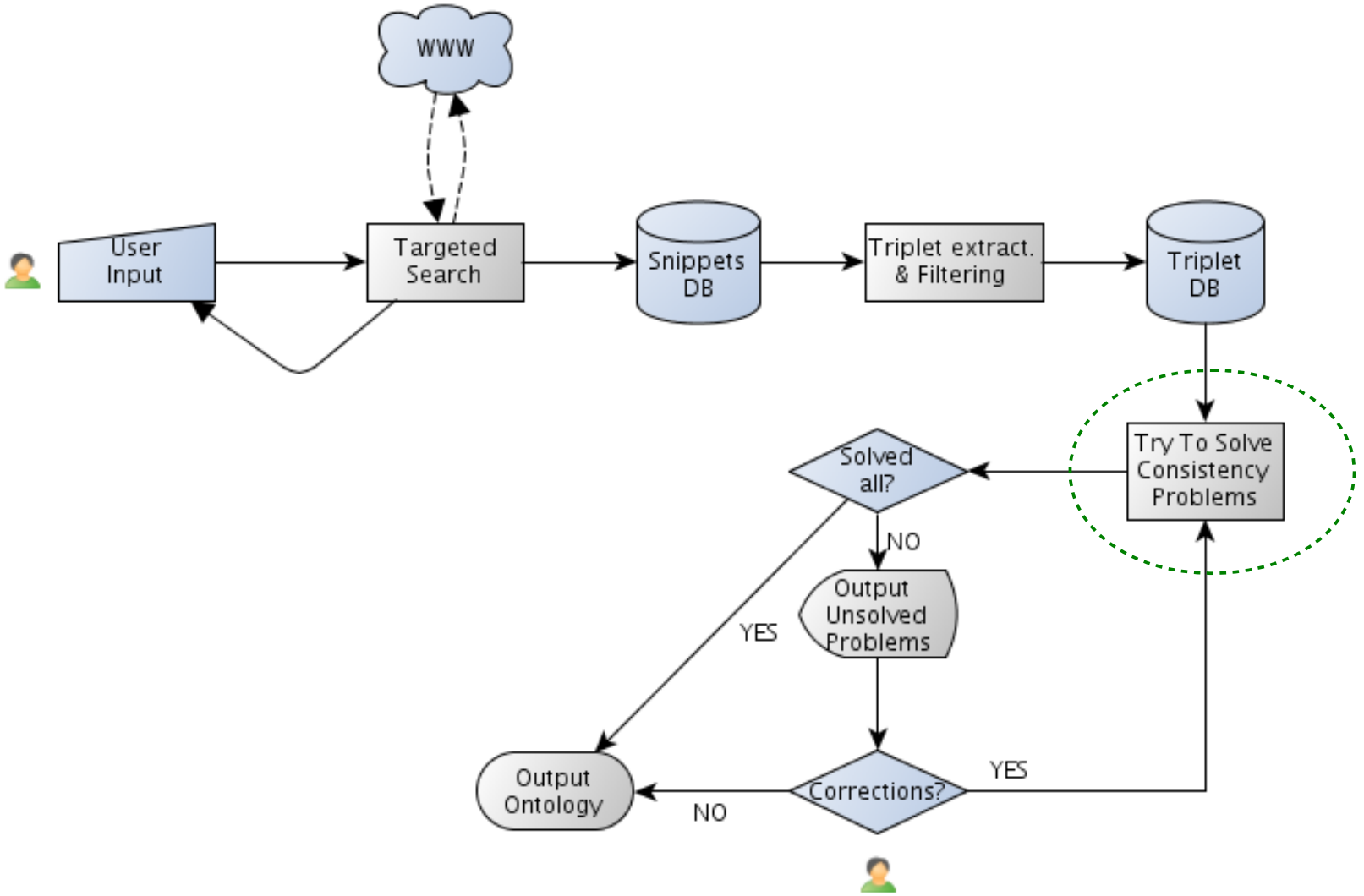
Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

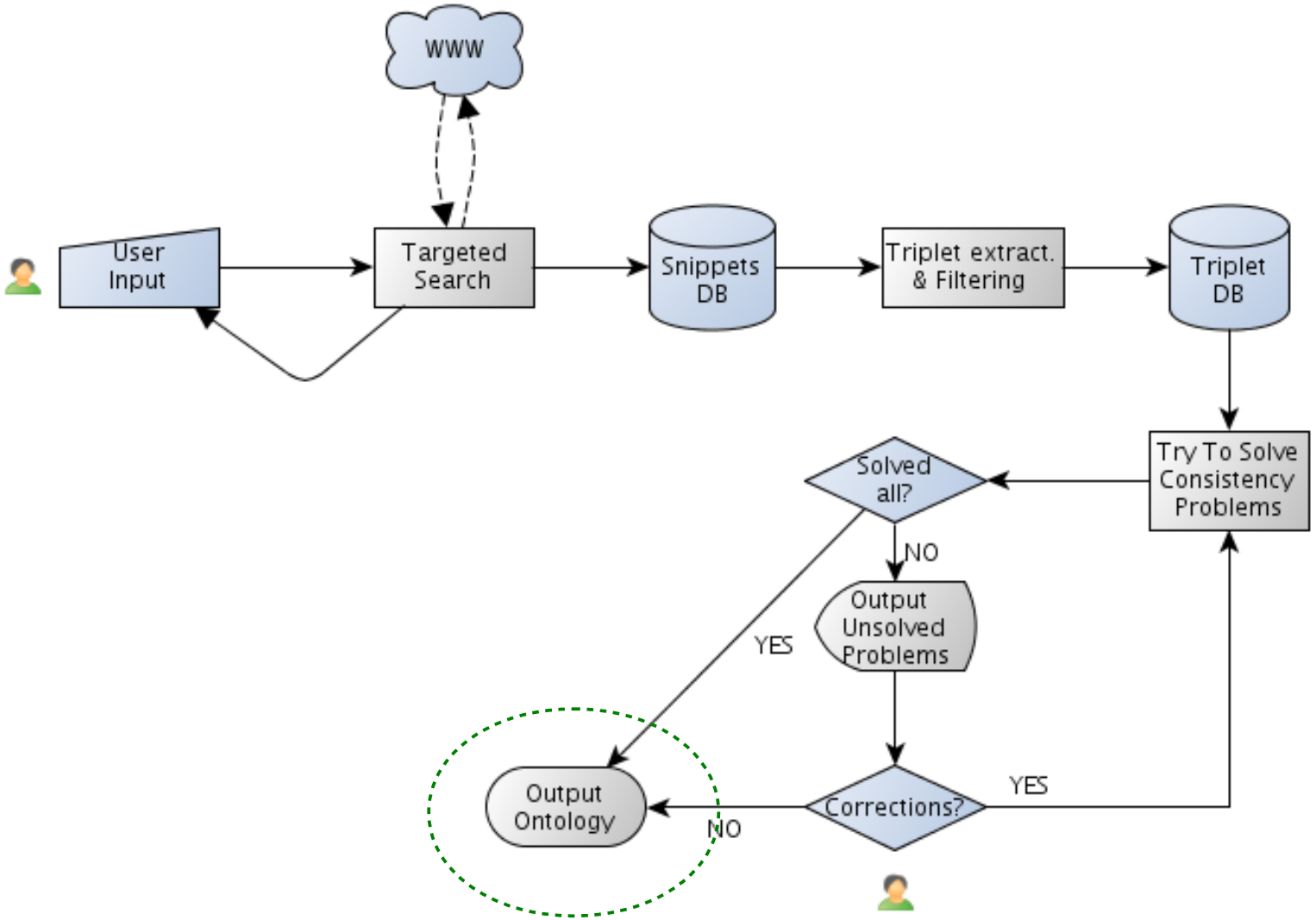
Overview


Input Fields
Search Company : Bristol Myers Squibb
acquired_direct2
Search
left
right
Type a name and press Enter/Search.
Edit Pane
Mar|9 7|9 ,|9 2012|9 ..|0 .|0 ZYMOGENETICS|1 ,|1 INC|1 (|0 acquired|3 by|3 Bristol|1 -|1 Myers|1 Squibb|1)|0 -|0 .|0 Atlanta|0 ,|0 GA|0 .|0 2007|9 to|0 2009|9 .|0 •|0 Exceeded|0 100|0 %|4 sales|0 and|0 management|0 objectives|0 ..|0 .|0
Retrain Model
Refresh
Add To Train Set Now
Annotated Snippets
ZymoGenetics was acquired by Bristol - Myers Squibb October 13 of last year for $9.75 per share , and Isilon Corp. was purchased by EMC Corp. on December 21 .. .
Edit
Ont entry
Mar 7 , 2012 .. . ZYMOGENETICS , INC (acquired by Bristol - Myers Squibb) - . Atlanta , GA . 2007 to 2009 . • Exceeded 100 % sales and management objectives .. .
Edit
Ont entry
Jan 11 , 2012.. . and another significant ' pearl ' , ZymoGenetics , was acquired . Fitch does recognize Bristol - Myers Squibb 's efforts since 2007 to strengthen the .. .
Edit
Ont entry
Mar 1 , 2011 .. . Previously , he was chief executive officer at ZymoGenetics Inc (acquired by BMS) , chief scientific officer at Seattle Genetics , and senior vice .. .
Edit
Ont entry
Jan 11 , 2012.. . and another significant ' pearl ' , ZymoGenetics , was acquired . Fitch does recognize Bristol - Myers Squibb 's efforts since 2007 to strengthen the .. .
Edit
Ont entry
Jul 19 , 2011 .. . In October 2010 , ZymoGenetics was acquired by Bristol - Myers Squibb . Posted by Tracy Sferra , Course Coordinator · Email ThisBlogThis !
BristolMyersSquibb acquired ZymogeneticsInc
Add Rellation
Add Company

Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

Overview


WWW
User Input
Targeted Search
Snippets DB
Triplet extract. & Filtering
Triplet DB
Try To Solve Consistency Problems
Solved all?
NO
YES
Output Unsolved Problems
Output Ontology
Corrections?
YES
NO

Overview

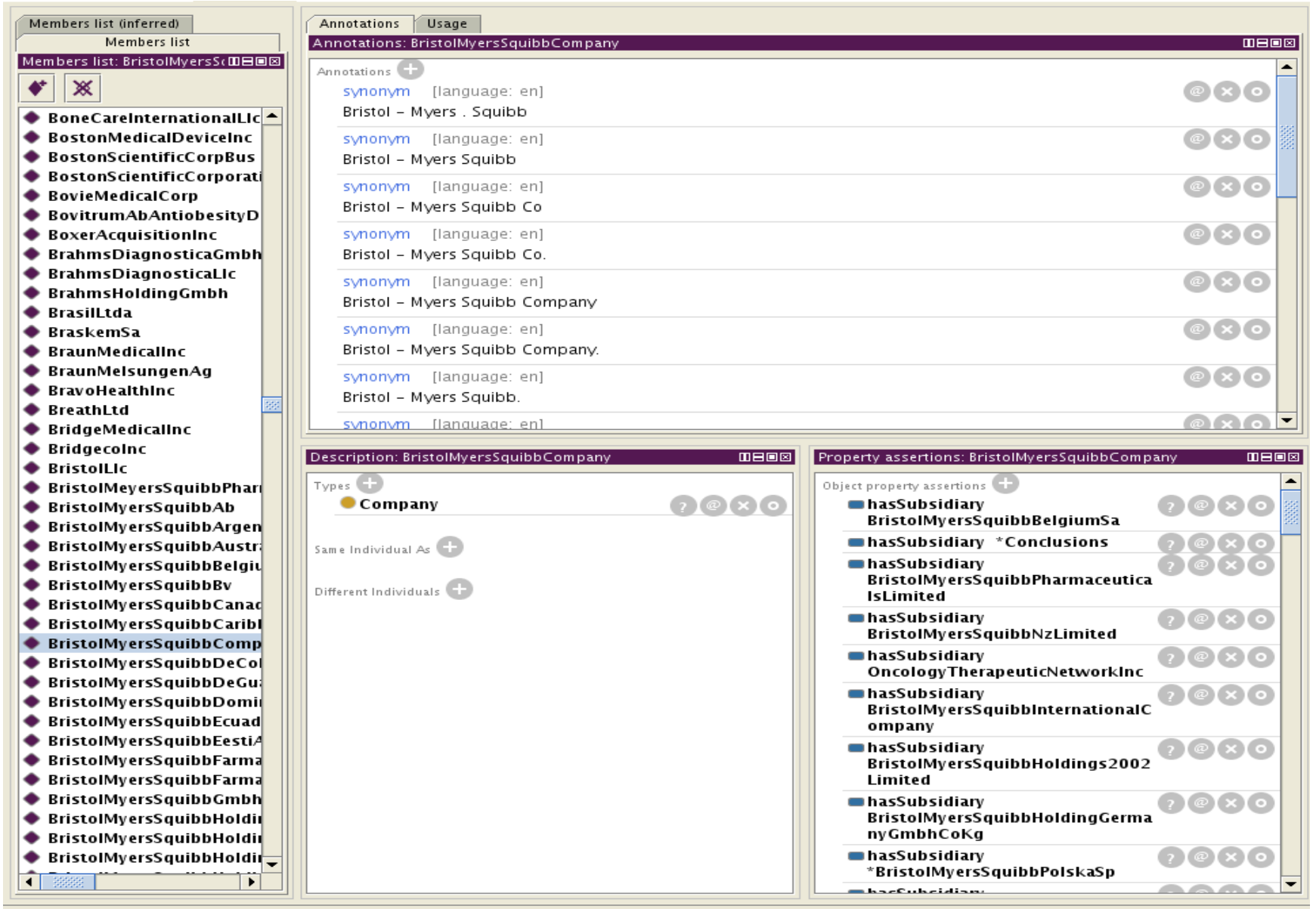


Figure: Screenshot of our ontology opened with Protege 4.2

Triplet Extraction

- Triplets are property assertions stored as 3-tuples
  - (subject, relation type, direct object)
- 3-step triplet extraction
  - Tag companies
  - Tag indicator words for relations
  - Match semantic patterns
- Company annotation
  - Conditional random fields (courtesy to Dr. Katrin Tomanek)
  - Trained on ~2000 snippets
  - 85% precision (5 folds cross validation)
  - *Bristol-Myers Squibb has acquired Adnexus Therapeutics, developer of a new class of ... Bristol-Myers Squibb also acquired Medarex Inc., a biotech company and a ...*

Triplet Extraction

- Triplets are property assertions stored as 3-tuples
  - (subject, relation type, direct object)
- 3-step triplet extraction
  - Tag companies
  - Tag indicator words for relations
  - Match semantic patterns
- Company annotation
  - Conditional random fields (courtesy to Dr. Katrin Tomanek)
  - Trained on ~2000 snippets
  - 85% precision (5 folds cross validation)
  - *Bristol-Myers Squibb has acquired Adnexus Therapeutics, developer of a new class of ... Bristol-Myers Squibb also acquired Medarex Inc., a biotech company and a ...*

Triplet Extraction (2)

- Tag indicator words
  - Exact formulations are prior defined by hand
  - Search expressions are a subset of the indicator expressions
  - *Bristol-Myers Squibb has acquired Adnexus Therapeutics, developer of a new class of ... Bristol-Myers Squibb also acquired Medarex Inc., a biotech company and a ...*
- Match semantic patterns
  - Inspired by *Hearst, 1992*
  - `company` $(otherTerms)\{0, 5\}$ `indW` $(otherTerms)\{0, 5\}$ `company`
  - Check the other terms
  - (`BristolMyersSquibb`, `acquired`, `AdnexusTherapeutics`), (`BristolMyersSquibb`, `acquired`, `MedarexInc`)

Triplet extraction (3)

- Whole snippets (like in the previous example)
- Contextually decomposed snippets
  - *“Merial, the Animal Health division of Sanofi, has acquired NewportLaboratories, a privately held company based in Worthington, Minnesota.”*
  - Context 0: Merial has acquired NewportLaboratories.
  - Context 1: Merial the Animal Health division of Sanofi
  - Context 2: Worthington, Minnesota
  - Context 3: NewportLaboratories, a privately held company based in Worthington
  - Courtesy to Elmar Hausmann

Ontology building

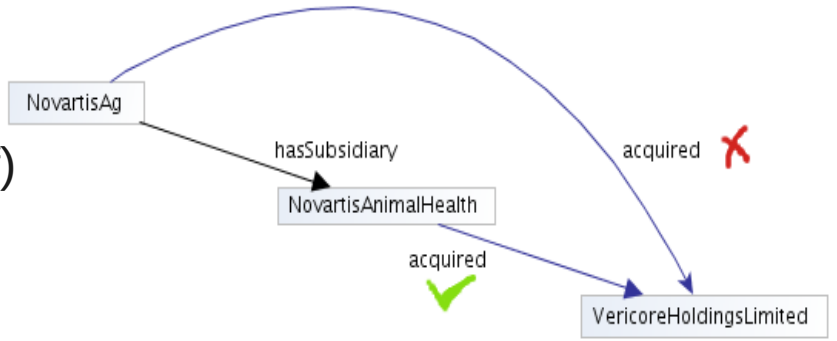
- Structure
  - 1 Class
  - 4 relation types `acquired`, `hasSubsidiary`, `hasMergerPart`, `holdsSharesBy`
- Main problems
  - Identity matching (Novartis = Novartis AG)
  - Injectivity violations, cycles, incorrect/incomplete information
  - Entity tagging problems

Identity matching

- Identification and merging of different name variaitions which refer to the same individual e.g.
  `(AbbottLaboratories, acquired, KosPharmaceuticalCo);`
  `(AbbottLabs, acquired, TapPharmaceuticals)`
- Assign synonyms & preferred terms
  - NovartisPharmaceuticalsCorporation: NovartisPharma, NovartisPharmaceuticals; NovartisAg: Novartis
  - Group names by root, get suffixes, assign preferred & synonyms
  - Based on the number of hits
  - Restriction: preferred terms always have a company suffix

Other Issues

- Named entity filtering
- Triplet filtering
  - $treeSize\_parent$ / $treeSize\_child$ < δ
  - $hits\_parent$ / $hits\_child$ < α
  - (JohnsonSonInc, acquired, BayerAg) and h_JohnsonSonInc = 35,600 , h_BayerAg = 2,790,000; st_JohnsonSonInc = 1 and st_BayerAg = 40
- Check for cycles
- Solve potential injectivity violations
  - Choose the most specific triplet (deepest leaf)
  - Prioritize companies with suffix
  - Occurence number of the triplets
    - *occNo(triplet_i)* / *occNo(triplet_j)* > 5



Evaluation Measures

- Standard precision, recall, F1 score
- The **triplet-wise recall** is a self made measure
  - important that a triplet is retrieved, but not that it is retrieved from every single snippet in which it occurs
  - *Apr 1 , 2009 ... Sanofi acquires Brazilian generics firm Medley for 500mm on Strategic Transactions , Apr-01-2009 . (snippet 1)*
    - *(*`Sanofi, acquired, Medley`*)*
  - *Dec 21 , 2009 ... Sanofi has acquired generic - drug makers HelvepharmAg of Switzerland, MedleySa of Brazil and LaboratoriosKendrickSa of Mexico,... (snippet 2)*
    - *(*`Sanofi, acquired, HelvepharmAg`*)*
  - *T-w recall =* ***2/3***
  - *Standard recall =* ***2/4***

Evaluation: Triplet extraction

We used 500 hand annotated snippets as gold standard.

	all	acquired	hasSubsidiary
Precision	84.97 %	82.93 %	87.21 %
Recall	58.10 %	61.26 %	59.06 %
F1 score	69.01 %	70.47 %	70.42%
Triplet-wise recall	86.21 %		

Table1 : Evaluation of our triplet extraction with contextual sentence decomposition of the snippets.

	all	acquired	hasSubsidiary
Precision	82.52 %	85.39 %	82.86 %
Recall	67.46 %	69.09 %	70.16 %
F1 score	74.23 %	76.38 %	75.98 %
Triplet-wise recall	92.96 %		

Table2 : Evaluation of our triplet extraction an whole snippets.

Co-occurence

- The **simple co-occurence** takes all pairs of entities in a snippet as positives.
  - (`NovartisAG, AlconInc`), unordered
- The **typed co-occurence** also takes into account the difficulty of relation type disambiguation and subject/object identification.
  - (`NovartisAG, holdsSharesBy, AlconInc`), (`NovartisAG, acquired, AlconInc`), (`NovartisAG, hasSubsidiary, AlconInc`), (`AlconInc, holdsSharesBy, NovartisAG`), (`AlconInc, acquired, NovartisAG`), (`AlconInc, hasSubsidiary, NovartisAG`)
- Lower limit for precision and a upper limit for recall

Measure	Simple co-occurence	Typed co-occurence
Recall	80.55 %	80.55 %
Precision	31.96 %	5.56 %
F1 score	45.77 %	10.40 %

Table3 : Co-occurence results on whole snippets

Ontology evaluation

Evaluated relations	Correct	Irrelevant	False
all	77.83 %	12.97 %	9.19 %
hasSubsidiary	93.9 %	3.66 %	2.44 %

Table4 : Ontology evaluation on 300 “child” companies owned by 4 parent companies

- **Correct**: Acquisitions, subsidiaries or divisions which could be found on official web sites
- **Irelevant**: entity tagger problems e.g. ***At****PhadiaUsInc, DoeIngallsInc****More*** or incomplete names
- **False**: No relevant evidence could be found

  Costs:

  - On average ~2 queries per ontology entry
  - Worst case scenario would be 8 queries per entry (2 test queries and 4 normalization queries)
  - 5$ for 125 triplets

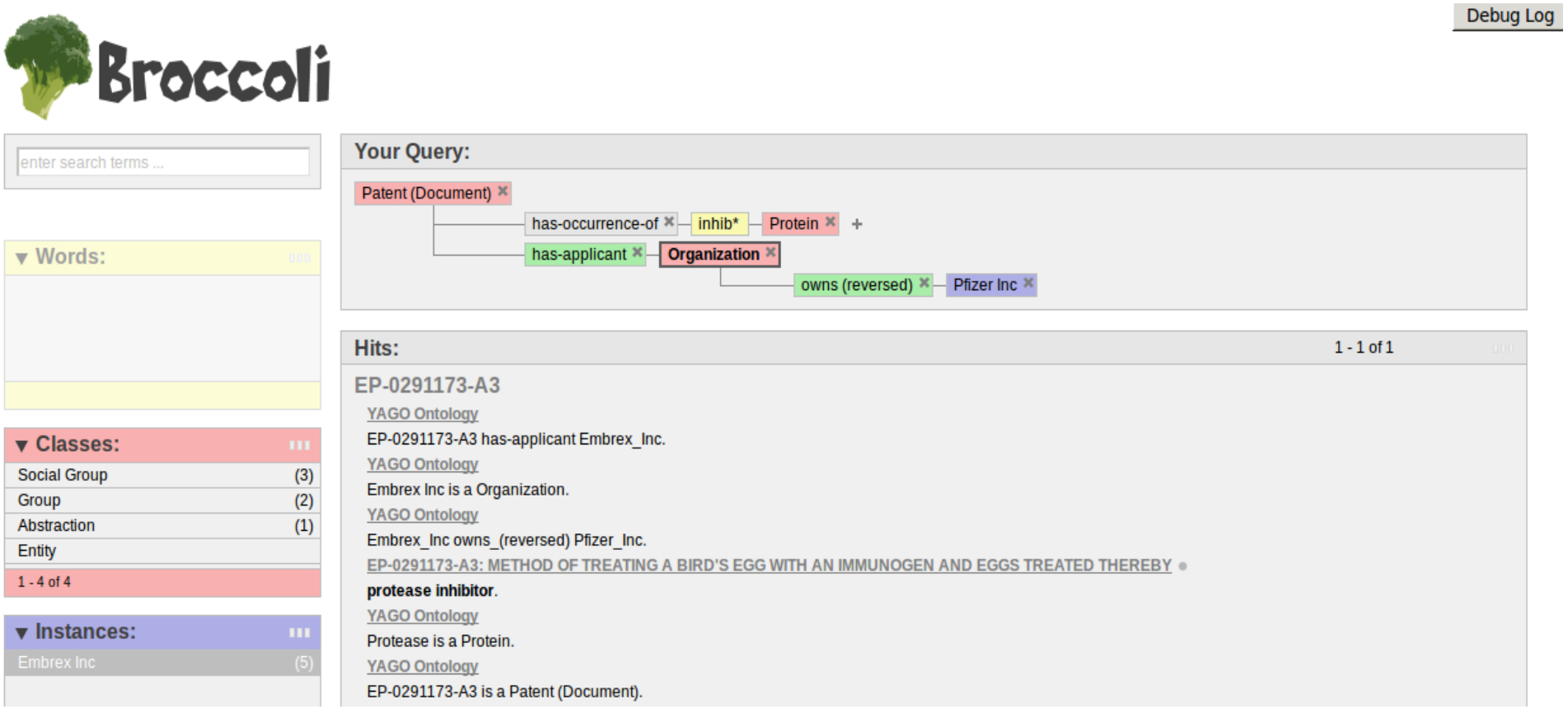
Patent Use Case

- Broccoli is a semantic search engine which combines full text search and ontology search
- Developed by Prof. Bast's group
- 1.5 Mio European patent documents from MAREC database
- Simple entity recognition, contextual decomposition
- Our company ontology and YAGO ontology

Example of possible query formulations:

- What patents are assigned to Company X and subsidiaries?
- On which proteins does Company X work?
- Which patents about protein inhibition are held by private persons? (instead of companies)

Patent Use Case (2)



http://filicudi.informatik.uni-freiburg.de:6222/BroccoliPatents/

Which regulation methods patented Company X?

Further Work

- Use a statistical model to evaluate the “trustworthiness”of a snippet (URL, writing style)
- Use a model to assign probabilities to already extracted triplets (“trustworthiness” of the snippet, occurrence number, hits number)
- Use more relations
- Solve temporal issues (i.e. reacquisitions)
- Completely automatize the process

Conclusion

- 92.9% triplet wise recall shows that the information in snippets is accessible to extraction tools despite their structure and poor grammar
- given the (overall) unreliable nature of the information on the Internet, filtering techniques are needed
- < 10 % false relations in our ontology shows that even simple tests and conflict solving approaches such as ours can be efficient

Thank you!

For your advices, time and attention.