Ontologies and Text Mining for Biomedicine

He Tan

Ontologies and Ontology Engineering 2011
Institutionen för datavetenskap

Outline

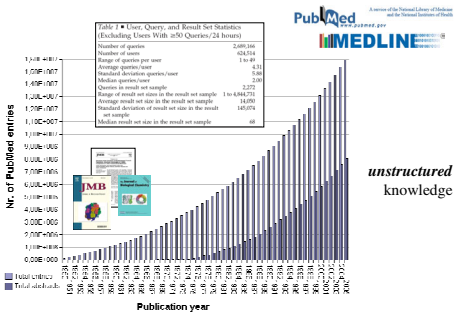
- Biomedical Text Mining
- Biomedical Ontologies
- TM systems using Ontologies
- Ontology motivated corpus construction

Biomedical Text Mining

- Individual gene study large scale analysis

- Genomics Biology: increasing number of genomes, sequences, proteins
- Large scale experimental results, e.g. Y2H, microarrays, etc.
- Large ***structured*** biological databases.

Biomedical Text Mining

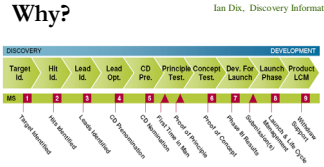


unstructured knowledge

A Day in the Life of PubMed: Analysis of a Typical Day's Query Log
Herskovic JR, et al. *J Am Med Inform Assoc.* 14(2): 212–220, 2007.

Why?

from the slide "Text Mining and Drug Discovery"
Ian Dix, Discovery Information AstraZeneca



- R&D:
  - Which proteins interact with metabolite X?
  - What are the reaction kinetics for canonical pathway Y?
  - Which compounds are associated with adverse event Z?
  - Which research groups are working on disease A?
  - What attributes are common to sets of biomarker genes?
  - What are the known associations between expressed genes and environmental factors?

Why?

- To make the right decision we need the relevant information:
  - Experimental data

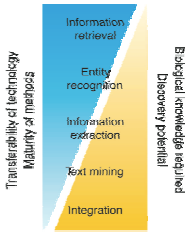
  AND
  - Contextual information
- Where is contextual information?
  - Unstructured text (~ 80%)
    internal documents + external research articles

Biomedical Text Mining

Linking genes to literature: text mining, information extraction, and retrieval applications for biology
Krallinger M. et. al *Genome Biology* 2008, 9(Suppl 2):S8

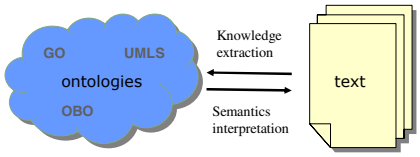
Disciplines

- **Information Retrieval (IR)**
  - finding the relevant documents
- **Entity Recognition (ER)**
  - identifying the entities
- **Information Extraction (IE)**
  - formalizing the facts
- **Text Mining (TM)**
  - finding nuggets in the literature
- **Integration**
  - combining text and biological data



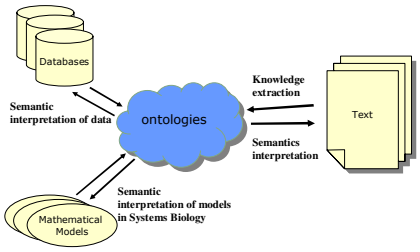
Literature mining for the biologist: from information retrieval to biological discovery
Jensen et al., *Nature Reviews Genetics*, 2006

Ontologies and Text Mining



Text mining and ontologies in biomedicine: Making sense of raw text
Spasic I et al., *Briefings in bioinformatics*, 6(3):239–251. 2005

Ontologies and Text Mining



Outline

- Biomedical Text Mining
- Biomedical Ontologies
- TM systems using Ontologies
- Ontology motivated corpus construction

The Gene Ontology (GO)

- The goal of the Gene Ontology Consortium (GOC) is to provide three ontologies of defined terms representing gene product properties.

34164 terms (100.0% defined)
- 20764 biological_process
- 2831 cellular_component
- 9010 molecular_function

May 16, 2011 at 13:38 Pacific time

relations,
- is_a,
- part_of
- regulates (positively_regulates, negatively_regulates)

Medical Subject Headings (MeSH)

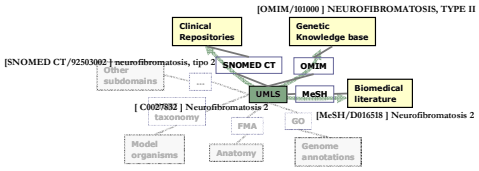
- The National Library of Medicine's (NLM) MeSH is the controlled vocabulary used for indexing articles for the MEDLINE® subset of PubMed.

  26,142 headings in 2011 MeSH
  arranged hierarchically by subject categories with more specific (narrower) terms arranged beneath broader terms.

  177,000 entry terms,
  e.g. "Vitamin C" is an entry term to "Ascorbic Acid."

UMLS Metathesaurus

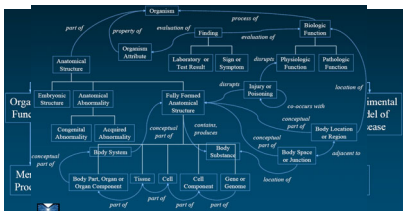
- Content
  - over 100 "source vocabularies "
  - 6M names, 1.5M concepts
  - 8M relations



UMLS Semantic Network

- Content
  - 135 high-level categories
  - 7000 relations among them

Concept [C0027832] Neurofibromatosis 2
Semantic Types Neoplastic Process [T191]



OBO Foundry

- The vision is that a core of these ontologies will be fully interoperable, *by virtue of a common design philosophy and implementation*, thereby enabling scientists and their instruments to communicate with minimum ambiguity.
- OBO Foundry Principles

1. The ontology must be *open* and available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed under the original name or with the same identifiers.
2. The ontology is in, or can be expressed in, a *common shared syntax*. This may be either the OBO syntax, extensions of this syntax, or OWL.
3. The ontologies possesses a *unique identifier space* within the OBO Foundry.
4. The ontology provider has procedures for identifying distinct successive *versions*.
5. The ontology has a clearly specified and clearly *delineated content*.
6. The ontologies include textual *definitions* for all terms.
7. The ontology uses relations which are unambiguously defined following the pattern of definitions laid down in the *OBO Relation Ontology*.
8. The ontology is *well documented*.
9. The ontology has a plurality of independent *users*.
10. The ontology will be developed *collaboratively* with other OBO Foundry members.

OBO Foundry

http://www.obofoundry.org/

Table 2 OBO Foundry ontologies (as of April 2007)

Ontology	Scope	URL	Custodians
Mature ontologies undergoing incremental reform			
Cell Ontology (CL)	Cell types from prokaryotic to mammalian	http://obofoundry.org/cgi-bin/detail.cgi?cell	Michael Ashburner, Jonathan Bard, Oliver Hofmann, Susa Rhee
Gene Ontology (GO)	Attributes of gene products in all organisms	http://www.geneontology.org	Gene Ontology Consortium
Foundational Model of Anatomy (FMA)	Structure of the mammalian and in particular the human body	http://fma.biostr.washington.edu	J.L.V. Mejino, Jr., Cornelius Rosse
Zebrafish Anatomical Ontology (ZAO)	Anatomical structures in *Danio rerio*	http://zfin.org/zf_info/anatomy/dict/sum.html	Melissa Haendel, Monte Westerfield
Mature ontologies still in need of thorough review			
Chemical Entities of Biological Interest (ChEBI)	Molecular entities which are products of nature or synthetic products used to intervene in the processes of living organisms	http://www.ebi.ac.uk/chebi	Paula Dematos, Rafael Alcantara
Disease Ontology (DO)	Types of human disease	http://diseaseontology.sf.net	Rex Chisholm
Plant Ontology (PO)	Flowering plant structure, growth and development stages	http://plantontology.org	Plant Ontology Consortium
Sequence Ontology (SO)	Features and properties of nucleic acid sequences	http://www.sequenceontology.org	Karen Eilbeck
Ontologies for which early versions exist			
Ontology for Clinical Investigations (OCI)	Clinical trials and related clinical studies	http://www.bioontology.org/wiki/index.php/OCI:Main_Page	OCI Working Group
Common Anatomy Reference Ontology (CARO)	Anatomical structures in all organisms	http://obofoundry.org/cgi-bin/detail.cgi?caro	Fabian Neuhaus, Melissa Haendel, David Sutherland
Environment Ontology	Habitats and associated spatial regions and sites	http://www.obofoundry.org/cgi-bin/detail.cgi?id=envo	Norman Morrison, Dawn Field
Ontology for Biomedical Investigations (OBI)	Design, protocol, instrumentation and analysis applied in biomedical investigations	http://obi.sf.net	OBI Working Group
Phenotypic Quality Ontology (PATO)	Qualities of biomedical entities	http://www.phenotypeontology.org	Michael Ashburner, Suzanna Lewis, Georgios Gkoutos
Protein Ontology (PRO)	Protein types and modifications classified on the basis of evolutionary relationships	http://pir.georgetown.edu/pro	Protein Ontology Consortium
Relation Ontology (RO)	Relations in biomedical ontologies	http://obofoundry.org/ro	Barry Smith, Chris Mungall
RNA Ontology (RnaO)	RNA three-dimensional structures, sequence alignments, and interactions	http://roc.bgsu.edu/	RNA Ontology Consortium

The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*, 25:1251 – 1255, 2007.

Outline

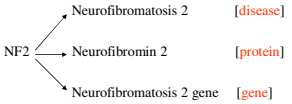
- Biomedical Text Mining
- Biomedical Ontologies
- TM systems using Ontologies
- Ontology motivated corpus construction

Biomedical Language

- Biomedical Language heavily use of domain specific terminology
  - e.g. *chemoattractant*, *fibroblasts*, *endocytosis*, *exocytosis*
- Short forms and abbreviations are often used
  - e.g. *vascular endothelial growth factor* (*VEGF*)
- Genes/proteins have often synonyms
  - e.g. *Thermoactinomyces candidus*
    vs. Thermoactinomyces vulgaris
- Orthographic variants
  - e.g. *TNFα, TNF-alpha* and *TNF alpha* (without hyphen)

Biomedical Language

- A name
  - general English term
  - may refer to a particular gene
  - may include homologues of this gene in other organisms
  - may denote an RNA, DNA, or the protein the gene encodes
  - may be restricted to a specific splice variant



ER – Ontologies as terminologies

- In practice, the distinction between ontological and terminological resources is somewhat arbitrary.
- Map ontological terms to free text
  - Features such as the position of a term in hierarchies and the semantic categorization of biomedical concepts can help disambiguate polysemous terms
  - Coverage
  - Term variation and management issues

IR – Search PubMed with MeSH

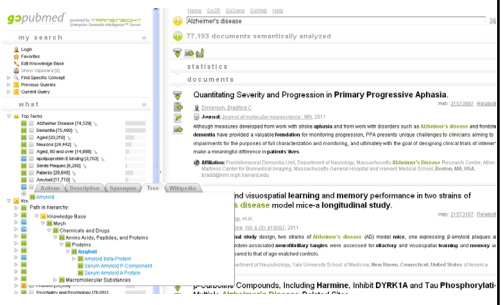
- Search Indexed for MEDLINE citations (90% of the PubMed database) using MeSH terms
  - MeSH term represents the major focus of the article
  - Assigned by professional human indexers at the NLM
- Limit searches to citations with the MeSH term
- Broaden/Narrow a search with the MeSH hierarchy
  - A search automatically include all articles which focus not only on the query term, but also focus on narrower terms
- Use subheadings to build complex and focused search strategies
  - Combine MeSH Terms

IR - GoPubMed

- It uses GO and MeSH to index search results,
  - Automatically map all ontology terms in GO and MeSH to the PubMed database.
  - Categorize the search results
  - Identify relevant terms
  - Summarize trends for a topic

GoPubMed: exploring PubMed with the Gene Ontology.
Doms A, Schroeder M, *Nucleic acids research*, 33: W783-W786, 2005.

IR - GoPubMed

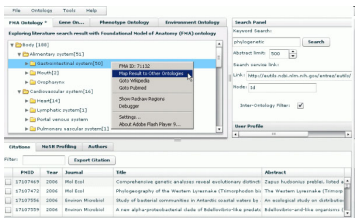


IR - PubOnto

- It uses ontologies from different perspective
  - Automatically mapping all ontology terms in GO, Foundational Model of Anatomy (FMA), Mammalian Phenotype Ontology, and Environment Ontology (ontologies from OBO foundary) to the full Medline database.
  - Inter-ontology filter mode shows intersecting articles between different ontologies.

Cross-Domain Neurobiology Data Integration and Exploration.
Xuan W, et al, *BMC Genomics*, 2010. (In press)

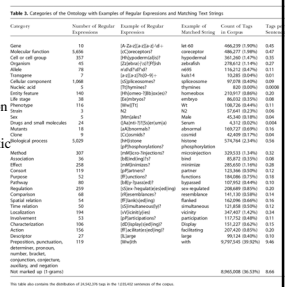
PubOnto



It provides the way to explore literature from different perspective.

IR/IE - Textpresso

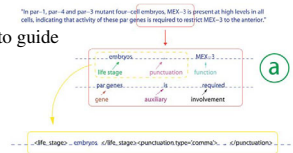
- Textpresso ontology
  - Categories
    - Biological entities
    - Characterize a biological entity or establish a relation between two of them
    - Auxiliary, used for semantic analysis of sentences.
  - Is automatically populated with 14,500 regular expressions (from a corpus of 3,307 journal articles)



Textpresso: an ontology-based information retrieval and extraction system for biological literature
Muller HM, er al, *PLoS Biol*. 2(11):e309, 2004

IR/IE - Textpresso

- Actively use ontology to guide and constrain analysis
  - Category searches



Assuming that many facts are expressed in one sentence, he would search for the categories "gene," "regulation," and "cell or cell group" in a sentence.

the question "What entities interact with 'daf-16' (a C. elegans gerontogene)?" can be answered by typing in the keyword "daf-16" and choosing the category "association."

IE - UMLS tools

- MetaMap
  - Map phrases in free text to concepts in the UMLS Metathesaurus
    - Several candidates
    - Candidate evaluation
    - Mapping score
- SemRep
  - Uses the Semantic Network to determine the relationship asserted between those concepts
    - Depends on the MetaMap results

IE – UMLS tools

- An example application

  The application of MetaMap and SemRep to an entire MEDLINE citation. The output provides a structured semantic overview of the contents of this citation.

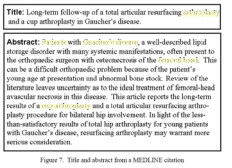


Figure 7. Title and abstract from a MEDLINE citation

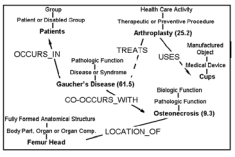


Figure 8. Semantic representation for the citation in Figure 7

IE/TM - PASTA

- PASTA Templates
  - 12 domain classes,
    e.g. *protein*, *residue*, *region*.
  - Object-oriented Template
    - template object
      a specific entity
    - a relation between objects,
      e.g *in_protein, in_species*
    - Scenario,
      e.g. a metabolic reaction
  - Slot filler: filling with information extracted from the text
    - Corpus - 1513 Medline abstract relevant to the study of protein structure.

```
<RESIDUE-134> :=
   NAME: ERGIN
   NO: 87
   SITE/FUNCTION: "catalytic"
                  "calcium-binding"
                  "active-site"
   SEC_STRUCT: "helical"
   QUAT_STRUCT: <not specified>
   REGION: "lid"
   INTERACTION: <not specified>

<IN_PROTEIN> :=
   RESIDUE: <RESIDUE-134>
   PROTEIN: <PROTEIN-2>

<IN_SPECIES> :=
   PROTEIN: <PROTEIN-2>
   SPECIES: <SPECIES-5>

<PROTEIN-2> :=
   NAME: "triacylglycerol lipase"

<SPECIES-5> :=
   NAME: "Pseudomonas cepacia"
```

Fig. 1. PASTA template examples

given the sentence *Ser154, Tyr167 and Lys171 are found at the active site*,

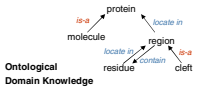
```
residue(e1), name(e1,"Ser154"),
residue(e2), name(e2,"Tyr167"),
residue(e3), name(e3,"Lys171"),
set(e4), set_member(e4, e1),
set_member(e4,e2), set_member(e4,e3),
find(e5), obj(e5,e4),
active_site(e6), at(e5,e6)
```

Protein structures and information extraction from biological texts: the PASTA system.
R. Gaizauskas, et al. *Bioinformatics*, 19(1): 135-143, 2003.

IE/TM - PASTA

- Discourse Processing
  - extract information from multiple sentences
  - make inferences using a limited predefined *domain ontology*.

S1 The three-dimensional structure of Endo H has been determined . . .
S5 A shallow curved cleft runs across the surface of the molecule from . . .
S6 This cleft contains the putative catalytic residue Asp130 . . .



S1: protein(e1), name(e1, "Endo H")
S5: cleft(e23), molecule(e25)
locate_in(e23,e25)
locate_in(e23, e1)

S6: cleft(e52),
residue(e61), name (e61, "Asp130")
contain(e52, e61)
locate_in(e61,e52)
locate_in(e61,e1)

i.e. Asp130 is in EndoH.

IE/TM – GenIE

Fig. 1. Linear IE architecture.

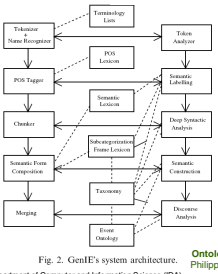


Fig. 2. GenIE's system architecture.

- GenIE (Genome Information Extraction)
  - A general IE framework.

Ontology-driven discourse analysis for information extraction
Philipp C, et al. *Data & Knowledge Engineering* 55(1): 59-83, 2005.

IE/TM - GenIE

- The lexical and knowledge sources
  - A lexicon of gene names for *Saccharomyces cerevisiae*
  - Semantic lexicon – 50 different linguistic variations of the term *yeast*, or *Saccharomyces cerevisiae* in 9000 Medline abstracts
  - POS Tagger – TreeTagger was trained on a manually annotated training corpus (600 SWISS-PROT function slots)
  - Sub-categorization frames for verbs
    - How to acquire the possible sub-categorization frames for all the relevant verbs?
      - ***binding*** events: *bind, binds, binding*
      - The frames are automatically acquired from a domain-specific corpus (Swiss-Prot function slots)

bind(subj(agent):protein,obj(patient):protein domain)

IE/TM - GenIE

- The lexical and knowledge sources
  - An ontology of biochemical events
    - Classification of biomchemical events
    - A taxonomy of biomchemical events
    - Relations between biochemical events
    - Ontology-driven approach to discourse analysis

Fig. 2. GenIE's system architecture.

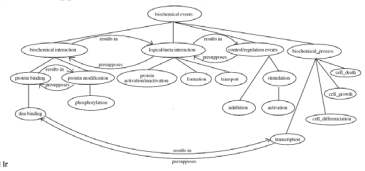


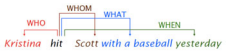
Fig. 7. Top level of the ontology O_{bio}.

Outline

- Biomedical Text Mining
- Biomedical Ontologies
- TM systems using Ontologies
- Ontology motivated corpus management

Semantic Role Labeling

- Semantic Role Labeling (SRL) is a process that, for each predicate in a sentence, indicates what semantic relations hold among the predicate and other sentence constituents that express the participants in the event.



- It is believed to play a key role in Information Extraction, Question Answering and Summarization.
- Large corpora annotated with semanticl roles
  - FrameNet
  - PropBank

PropBank

- Penn TreeBank → PropBank
  - Add a semantic layer on Penn TreeBank
  - Define a set of semantic roles for each verb
  - VerbNet project maps PropBank verb types to their corresponding Levin classes

hit.01 "strike"
- A0: agent, hitter; A1: thing hit;
  A2: instrument, thing hit by or with

[A0 *Kristina*] ***hit*** [A1 *Scott*] [A2 *with a baseball*] *yesterday*. AM-TMP *Time*

FrameNet

- Sentences from the British National Corpus
- Method of building FrameNet
  - collects and analyzes the corpus attestations of target words with semantic overlapping.
  - The attestations are divided into semantic groups, and then these small groups are combined into frames.

[Agent *Kristina*] ***hit*** [Target *Scott*] [Instrument *with a baseball*] [Time *yesterday*].

FrameNet vs. PropBank

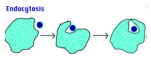
- FrameNet includes semantic analysis of all major parts of speech, not only verb.
- PropBank makes reference to specific tree nodes of TreeBank's syntactic parses of the corpus data.
- In FrameNet, different lexical units within the same Frame will have consistent uses of semantic roles.

Domain-Specific Corpus

- As with other technologies in natural language processing (NLP), researchers have experienced the difficulties of adapting SRL systems to a new domain, different than the domain used to develop and train the system.
- Biomedical text considerably differs from the text in these corpus, both in the style of the written text and the predicates involved.

Biomedical Language

- Predicates in biomedical text often prefers nominalizations, gerunds and relational nouns
  - *e.g. interaction, association, binding, transcription*
- Domain specific predicates are absent from both the FrameNet and PropBank data
  - e.g *endocytosis*, *exocytosis* and *translocate*
- Predicates have been used in biomedical documents with different semantic senses and require different number of semantic roles compared to FrameNet and PropBank data
  - e.g *block*, *generate* and *transform*,



Difficulties of building frame lexicon

- How to discover and define semantic frames together with associated semantic roles within the domain?
- How to collect and group domain-specific predicates to each semantic frame?
- How to select example sentences from publication databases, such as the PubMed/MEDLINE database containing over 20 million articles?

Biological procedure ontology of GO

GO:0008150 : biological_process [353603 gene products]
- GO:0051234 : establishment of localization [40515 gene products]
  - GO:0045184 : establishment of protein localization [8102 gene products]
    - **GO:0015031 : protein transport** [7966 gene products]
      - GO:0071828 : apolipoprotein E recycling [0 gene products]
      - GO:0043213 : bacteriocin transport [11 gene products]
      - GO:0072321 : chaperone-mediated protein transport [4 gene products]
      - GO:0034436 : glycoprotein transport [3 gene products]
      - GO:0006886 : intracellular protein transport [4290 gene products]
      - GO:0033571 : lactoferrin transport [0 gene products]
      - GO:0042953 : lipoprotein transport [58 gene products]
      - GO:0051224 : negative regulation of protein transport [240 gene products]
      - GO:0051222 : positive regulation of protein transport [424 gene products]
      - GO:0017038 : protein import [1420 gene products]
      - GO:0009306 : protein secretion [1500 gene products]
      - GO:0071806 : protein transmembrane transport [604 gene products]
      - GO:0072322 : protein transport across periplasmic space [0 gene products]
      - GO:0071693 : protein transport within extracellular region [0 gene products]
      - GO:0051223 : regulation of protein transport [795 gene products]
      - GO:0033572 : transferrin transport [12 gene products]
      - GO:0051808 : translocation of peptides or proteins into other organism involved in symbiotic interaction [0 gene products]
  - GO:0006810 : transport [40099 gene products]
    - **GO:0015031 : protein transport** [7966 gene products]
      - GO:0071828 : apolipoprotein E recycling [0 gene products]
      - GO:0043213 : bacteriocin transport [11 gene products]

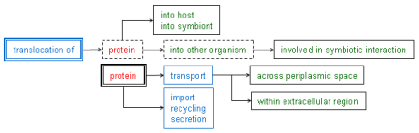
"the directed movement of proteins into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore".

177 descendant classes.

581 class names and synonyms

Underlying compositional structures in ontological terms

9 direct subclasses of *protein transport*



The possible predicates *translocation, import, recycling, secretion and transport*

The more complex expressions, e.g. "*translocation of peptides or proteins into other organism involved in symbiotic interaction*" (GO:0051808), express participants involved in the event, i.e. **the entity** (*peptides or proteins*), **destination** (*into other organism*) and **condition** *(involved in symbiotic interaction)* of the event.

Aspects of the method

- The structure and semantics of domain knowledge in ontologies constrain the frame semantics analysis, i.e. decide the coverage of semantic frames and the relations between them;
- Ontological terms can comprehensively describe the characteristics of events/scenarios in the domain, so domain-specific semantic roles can be determined based on terms;
- Ontological terms provide a list of domain specific predicates, so the semantic sense of the predicates in the domain are determined;
- The collection and selection of example sentences can be based on knowledge-based search engine for biomedical text.

"*Protein Transport*" Frame

FEs	definition
Transport_Entity (TE)	Protein or protein complex which is undergoing the motion event into, out of or within a cell, or between cells, or within a multicellular organism.
Transport_Origin (TO)	The organelle, cell, tissue or gland from which the Transport_Entity moves to a different location.
Transport_Destination (TDS)	The organelle, cell, tissue or gland to which the Transport_Entity moves from a different location.
Transport_Condition (TC)	The event, substance, organelle or chemical environment which positively or negatively directly influences or is influenced by, the motion event. The substance organelle does not necessarily move with the Transport_Entity
Transport_Location (TL)	The organelle, cell, tissue or gland where the motion event takes place when the origin and the destination are the same or when origin or destination is not specified.
Transport_Path (TP)	The substance or orgnanelle which helps the entity to move from the Transport_Origin to the Transport_Destination, sometimes by connecting the two locations, without itself undergoing translocation
Transport_Transporter (TT)	The substance, organelle or cell crucial to the motion event, that moves along with the Transport_Entity, taking it from the Transport_Origin to the Transport_Destination.
Transport_Direction	The direction in which the motion event is taking place with re-

#T	578	80	159	95	41	25	6	2	1
FEs	TE	TO	TDS	TC	TL	TP	TT	TDR	TA

Table 3: The number of the GO terms that describe the frame elements

the head of GO terms	delivery (1), egress (2), establishment of ... localization (19), exit (2), export (20), import (88), recycling (2), release (1), secretion (226), sorting (4), targeting (68), trafficking (1), translocation (76), transport (100), uptake (5)
LUs	delivery.n, deliver.v, egress.n, establishment of ... localization.n, exit.n, exit.v, export.n, export.v, import.n, import.v, recycling.n recycle.v, release.n, release.v, secretion.n, secrete.v, sort.v, sorting.n, target.v, targeting.n, translocation.n, translocate.v, transport.v, transport.n, trafficking.n, uptake.n

Table 4: The lexical units

[L.pneumophila$_{Transport_Origin}$[NP_Eco]] [**translocate**$_{predicate}$] [more than 100 effector proteins$_{Transport_Entity}$[NP_Obj]] [into host cytoplasm$_{Transport_Destination}$[PP[into].Dep]] [using Dot/Icm T4BSS$_{Transport_Path}$[VPing.Dep]] [modulating host cellular functions$_{Transport_Condition}$[VPing.Dep]] to establish a replicative niche within host cells. (PMID: 20949065)

Table 5: An example sentence: the three layers of annotations are given as FE|PT.GF.

The Frame Lexicon

- First work on ontology driven corpus management
- Released the corpus covering "*protein transport*" event, http://www.ida.liu.se/~hetan/bio-onto-frame-corpus/
- We aim to extend the corpus to cover other biological events.
  - GO ontologies, other ontologies (e.g. pathway ontologies)
- The identification of frames and the relations between frames are needed to be investigated.
- We will study the definition of Semantic Type (ST) in the domain corpus and their mappings to classes in top domain ontologies
  - ST: "Sentient" defined for the semantic role "Cognizer" in the frame "Cogitation".

Thanks