Aiding the Data Integration in Medicinal Settings by Means of Semantic Technologies

Vit Novacek¹ Loredana Laera² Siegfried Handschuh¹

¹Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
email: {first_name.last_name}@deri.org

²Department of Computer Science
University of Liverpool, UK
email: lori@csc.liv.ac.uk

ESTC2007/MSWFB, Vienna, Austria — May 31, 2007
Outline

1 Introduction
   - Introduction to the Problem
   - Motivation

2 Architecture of the Framework
   - Broader Context
   - Ontology Integration

3 Using the Framework
   - Usage Example
   - Evaluation

4 Selected Life Science Use Cases

5 Conclusions and Future Work
   - Conclusions
   - Future Work

V. Novacek, L. Laera, S. Handschuh

Aiding the Data Integration in Medicinal Settings...
Introduction to the Problem

- context of our work – ontology evolution (Knowledge Web EU NoE)
- development of a simple methodology of ontology lifecycle scenario
- implementation of a respective framework, unifying all phases of the ontology lifecycle – DINO (Dynamics, INtegration or Data, INtensive; Ontologies)
- universal application, however, life-sciences and bio-medicine are our primary concern (due explicit needs for semantic solution and other domain specialties)
- critical and non-trivial task not covered by the state-of-the-art – ontology integration
Desired Features of the Integration

1. process *new knowledge* *semi-automatically* in dynamic domains
2. automatically compare the *new* and *current* knowledge
3. resolve and/or mark possible major *inconsistencies* between the new and current knowledge
4. automatically order the *new knowledge* according to user-defined preferences
5. transform the *new knowledge* into a form of *sorted suggestions* in simple *natural language*, alleviating human efforts in the task of the final *incorporation* of new knowledge
Why Healthcare?

- explicit need for semantic solutions:
  - data stored in unstructured or disparate repositories, multiple formats hampering interoperability
  - cannot be queried to the full potential efficiently within traditional solutions (e.g. databases)

- dynamic nature of the domain

- new, sometimes even critical, knowledge continually appears and has to be efficiently processed and integrated

- emphasis on easy-to-use solutions, since medical experts are generally not experts in data or ontology engineering
Outline

1 Introduction
   - Introduction to the Problem
   - Motivation

2 Architecture of the Framework
   - Broader Context
   - Ontology Integration

3 Using the Framework
   - Usage Example
   - Evaluation

4 Selected Life Science Use Cases

5 Conclusions and Future Work
   - Conclusions
   - Future Work
Ontology Lifecycle Scenario

V. Novacek, L. Laera, S. Handschuh

Aiding the Data Integration in Medicinal Settings...
Ontology Integration

DINO Integration Scheme

Ontology Learning Wrapper

Ontology A/N Wrapper

Ontology CD Portal Interface

Ontology R/M Wrapper

NL Suggestions Generator

Domain Resources

Ontology Diff Wrapper

User Preferences

Triple Sorter
...while cerebellar astrocytoma is usually discovered by means of CT... using a diagnostic procedure of scanning... GVHD, an immune dysfunction... GVHD, a disease being a type of dysfunction...
Learned Ontology

```xml
<owl:ObjectProperty rdf:ID="discovered-by"/>
<owl:Thing rdf:ID="CT"/>
<owl:Thing rdf:ID="cerebellar-astrocytoma">
  <discovered-by rdf:resource="#CT"/>
</owl:Thing>
<owl:Class rdf:ID="diagnostic-procedure"/>
<owl:Class rdf:ID="immune-dysfunction"/>
<owl:Class rdf:ID="dysfunction"/>
<owl:Class rdf:ID="scanning">
  <rdfs:subClassOf rdf:resource="#diagnostic-procedure"/>
</owl:Class>
<owl:Class rdf:ID="GVHD"/>
<owl:Class rdf:ID="disease">
  <rdfs:subClassOf rdf:resource="#dysfunction"/>
</owl:Class>
```
Master Ontology

```xml
<owl:ObjectProperty rdf:ID="InstrumentalProperty"/>
<owl:ObjectProperty rdf:ID="DiscoveredUsing">
  <rdfs:subPropertyOf rdf:resource="#InstrumentalProperty"/>
  <rdfs:range rdf:resource="#Manifestation"/>
  <rdfs:domain rdf:resource="#DiagnosisProcedure"/>
</owl:ObjectProperty>
<owl:Class rdf:ID="Manifestation"/>
<owl:Class rdf:ID="Procedure"/>
<owl:Class rdf:ID="DiagnosisProcedure">
  <rdfs:subClassOf rdf:resource="#Procedure"/>
</owl:Class>
<owl:Class rdf:ID="SoftTissueCytoma"/>
<owl:Class rdf:ID="AstroCytoma">
  <rdfs:subClassOf rdf:resource="#SoftTissueCytoma"/>
</owl:Class>
<owl:Class rdf:ID="Disease"/>
<owl:Class rdf:ID="Dysfunction">
  <rdfs:subClassOf rdf:resource="#Disease"/>
</owl:Class>
```
Agreed Mapping

<owl:ObjectProperty rdf:ID="DiscoveredUsing">
  <owl:equivalentProperty rdf:resource="#discovered-by"/>
</owl:ObjectProperty>
<AstroCytoma rdf:ID="cerebellar-astrocytoma"/>
<owl:Class rdf:ID="DiagnosisProcedure">
  <owl:equivalentClass rdf:resource="#diagnostic-procedure"/>
</owl:Class>
<owl:Class rdf:ID="immune-dysfunction">
  <owl:subClassOf rdf:resource="#Dysfunction"/>
</owl:Class>
<owl:Class rdf:ID="Dysfunction">
  <owl:equivalentClass rdf:resource="#dysfunction"/>
</owl:Class>
Refining the Merge by Inference

- inconsistency resolution:
  - disease and dysfunction are said to be subclasses of each other
  - the learned inconsistent assertion (disease < dysfunction) is therefore removed by default

- learned knowledge augmentation:
  - using range and domain of the DiscoveredUsing property in the master ontology, we can infer that:
    - cerebellar astrocytoma is an instance of Manifestation
    - CT is an instance of DiagnosisProcedure
Resulting Suggestions

Config:\  \quad w_c = 1.0 \quad w_r = 1.0 \quad \rho = 0.2 \quad t = 5
Pos  ::  Scanning discover cytoma
Neg  ::  subclass disease dysfunction

----------

+0.667::  CEREBELLAR ASTROCYTOMA is a new instance of ASTROCYTOMA.
+0.667::  CEREBELLAR ASTROCYTOMA is a new instance of MANIFESTATION.
+0.389::  CT is a new instance of DIAGNOSIS PROCEDURE.
+0.333::  GVHD is a new instance of IMMUNE DYSFUNCTION.
-0.444::  A new class SCANNING is a sub-class of DIAGNOSIS PROCEDURE.
-0.667::  CEREBELLAR ASTROCYTOMA is DISCOVERED USING CT.
-0.833::  A new class IMMUNE DYSFUNCTION is a sub-class of DYSFUNCTION.
Preliminary Evaluation and Current State

- **preliminary evaluation:**
  - sorting places 80.7% of triples correctly compared to an order given by a human user (on small artificial sample)
  - negotiation component has been evaluated using the Ontology Alignment Evaluation Initiative test suite, preliminary results promising

- **current state:**
  - testing, tuning and debugging of the full implementation of the framework presented in the paper
  - combination with new concept in MarcOnt Portal – MarcOnt Portal services – and collaborative Protégé initiated within implementation of the whole lifecycle framework
Outline

1 Introduction
   - Introduction to the Problem
   - Motivation

2 Architecture of the Framework
   - Broader Context
   - Ontology Integration

3 Using the Framework
   - Usage Example
   - Evaluation

4 Selected Life Science Use Cases

5 Conclusions and Future Work
   - Conclusions
   - Future Work
Longitudinal Electronic Health Record

- **needs:**
  - platforms supporting creation and management of long-term EHR
  - integration of different data sources
  - population of common conceptual structure (once it has been created)
  - efficient and expressive querying

- **solutions:**
  - ontologies bound to patient data
  - means for dynamic population of patient records from diverse resources (using learning, alignment and integration)
  - querying for free – using the state of the art OWL reasoning tools
Epidemiological Registries

- **needs:**
  - *population-wise* health records
  - *extension* of the needs in longitudinal EHR
  - *integration* and *selection* of the knowledge in patient records

- **solutions:**
  - *merging* of patient records, *filtering* the knowledge within the integration
  - *population* of an epidemiological ontology
  - again, querying using the *state of the art* OWL reasoning tools, adding *symbolic* dimension to the traditional *statistic* processing
Public Health Surveillance

**needs:**
- ongoing collection, analysis and dissemination of health-related data in order to facilitate a public health action
- needs more or less the same as in the previous case
- however, emphasis on efficient dynamic processing of new data

**solutions:**
- generic ontology integration and population services
- explicit support for efficient dynamic integration of new knowledge from textual resources (by ontology learning)
Management of Clinical Trials

- **needs:**
  - electronic representation of clinical trials data
  - heterogeneity and integration problems (usually, several different institutions involved)
  - cost-effective querying demanded

- **solutions:**
  - ontologies developed and/or mediated using the DINO framework
  - querying of different clinical trial data straightforward
Genomics and Proteomics Research

- **needs:**
  - bridging the research and clinical practice
  - integrate specific knowledge e.g. in GO or UMLS – medical controlled dictionaries
  - efficient symbolic querying

- **solutions:**
  - aiding semi-automatic ontology development
  - data mediation using ontology integration
  - even for not very well specified data, the mechanism of sorted suggestions generation can reduce the efforts in merging the knowledge
Outline

1. Introduction
   - Introduction to the Problem
   - Motivation

2. Architecture of the Framework
   - Broader Context
   - Ontology Integration

3. Using the Framework
   - Usage Example
   - Evaluation

4. Selected Life Science Use Cases

5. Conclusions and Future Work
   - Conclusions
   - Future Work
Conclusions

- DINO – a mechanism for dynamic ontology integration – introduced
- based on ontology learning, meaning negotiation, merging, refinement and generation of suggestions in natural language, sorted according to user-defined relevance
- preliminary evaluation, current state of the implementation reported
- importance of the ontology lifecycle framework for integration of more general healthcare data described using realistic use cases
Future Work

- improve the natural language generation mechanism
- finish the basic testing and tuning of the platform
- include the DINO integration into the broader context of the lifecycle platform implementation
- employ and test the whole framework in realistic settings in a healthcare industry, possibly in line with the presented use cases
- incorporate the feedback and challenges identified in the realistic evaluation within further improvement of the framework