REGULAR PATHS IN SPARQL: QUERYING THE NCI THESAURUS

Landon T. Detwiler[1], Dan Suciu, PhD[3], James F. Brinkley, MD, PhD[1,2,3]

University of Washington Structural Informatics Group
Departments of Biological Structure[1], Medical Education and Biomedical Informatics[2], and Computer Science and Engineering[3]

Introduction

- OWL (Web Ontology Language)
  - Basis in description logics
- Problem
  - Logical constructs have a cognitive cost
- Solution
  - Intuitive OWL views
    - SparQL based
    - Gleen – regular path enhancements
- Examples
  - NCI Thesaurus
    - Widely used
    - Exhibits common OWL representational patterns

Background

OWL and SparQL

OWL/RDF Triples

Subject	Predicate	Object
A	part	B
B	part	C
C	part	D
D	contains	E

SparQL Triple Patterns

```
SELECT ?r1 ?r4
FROM <htp://example.com/ex>
WHERE
{
  ?r1  part  ?r2 .
  ?r2  part  ?r3 .
  ?r3  contains  ?r4 .
}
```

Subject	Predicate	Object
A	part	B
B	part	C
C	part	D
D	contains	E

SparQL Triple Patterns

```
SELECT ?r1 ?r4
FROM <htp://example.com/ex>
WHERE
{
  ?r1 part ?r2 .
  ?r2 part ?r3 .
  ?r3 contains ?r4 .
}
```

Subject	Predicate	Object
A	part	B
B	part	C
C	part	D
D	contains	E

SparQL Triple Pattern Conjunction

```
SELECT ?r1 ?r4
FROM <htp://example.com/ex>
WHERE
{
   ?r1  part  ?r2 .
   ?r2  part  ?r3 .
   ?r3  contains  ?r4 .
}
```

Subject	Predicate	Object
A	part	B
B	part	C
C	part	D
D	contains	E

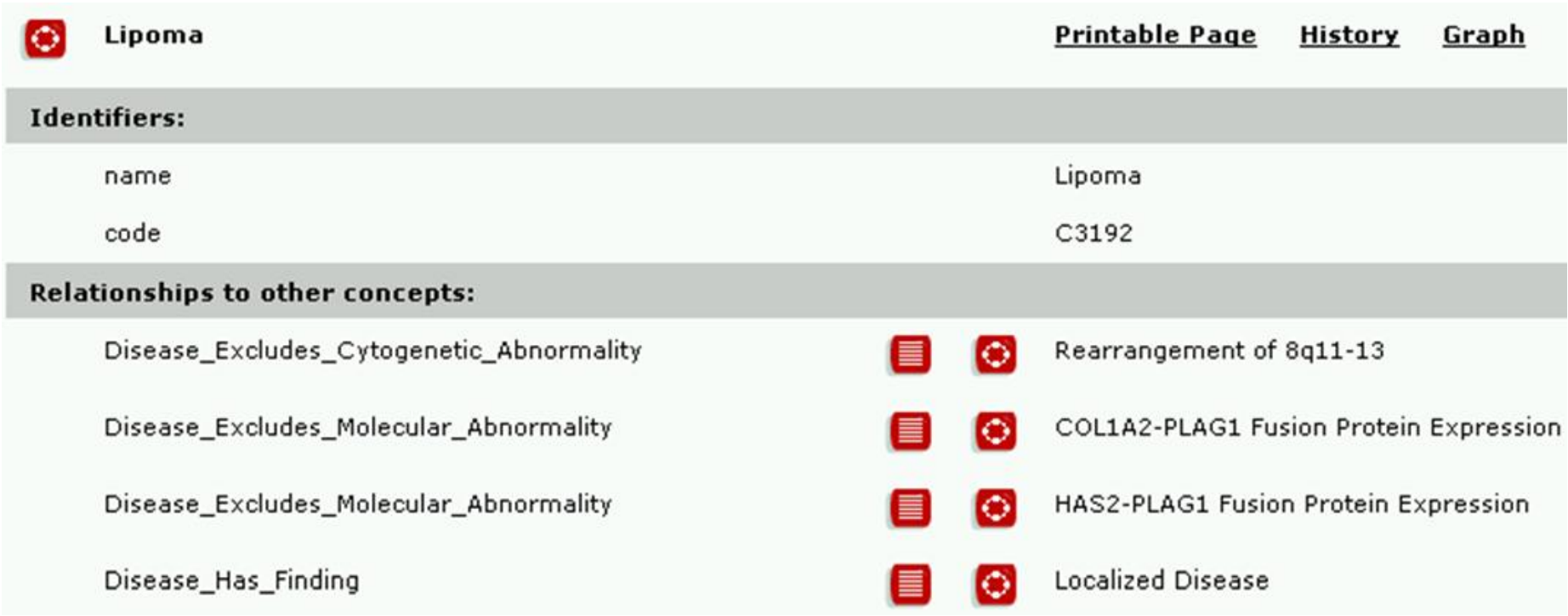
$?r1 \leftarrow B$
$?r2 \leftarrow C$
$?r3 \leftarrow D$
$?r4 \leftarrow E$

r1	r4
B	E

Problem

Querying for “direct” relationships in OWL

NCIt Browser View: Lipoma



Lipoma — Printable Page | History | Graph

Identifiers:	
name	Lipoma
code	C3192
Relationships to other concepts:	
Disease_Excludes_Cytogenetic_Abnormality	Rearrangement of 8q11-13
Disease_Excludes_Molecular_Abnormality	COL1A2-PLAG1 Fusion Protein Expression
Disease_Excludes_Molecular_Abnormality	HAS2-PLAG1 Fusion Protein Expression
Disease_Has_Finding	Localized Disease

Browser View Triples

Subject	Predicate	Object
Lipoma	label	“Lipoma”
Lipoma	code	“C3192”
Lipoma	Disease_Excludes_Cytogenetic_Abnormality	Rearrangement_of_8q11-13
Lipoma	Disease_Excludes_Molecular_Abnormality	COL1A2-PLAG1_Fusion_Protein_Expression
Lipoma	Disease_Excludes_Molecular_Abnormality	HAS2-PLAG1_Fusion_Protein_Expression
Lipoma	Disease_Has_Finding	Localized_Disease

SparQL "Directly Related" Query

Subject	Predicate	Object
Lipoma	label	"Lipoma"
Lipoma	code	"C3192"
Lipoma	Disease_Excludes_Cytogenetic_Abnormality	Rearrangement_of_8q11-13
Lipoma	Disease_Excludes_Molecular_Abnormality	COL1A2-PLAG1_Fusion_Protein_Expression
Lipoma	Disease_Excludes_Molecular_Abnormality	HAS2-PLAG1_Fusion_Protein_Expression
Lipoma	Disease_Has_Finding	Localized_Disease

Lipoma ?predicate ?object .

Expected Query Results

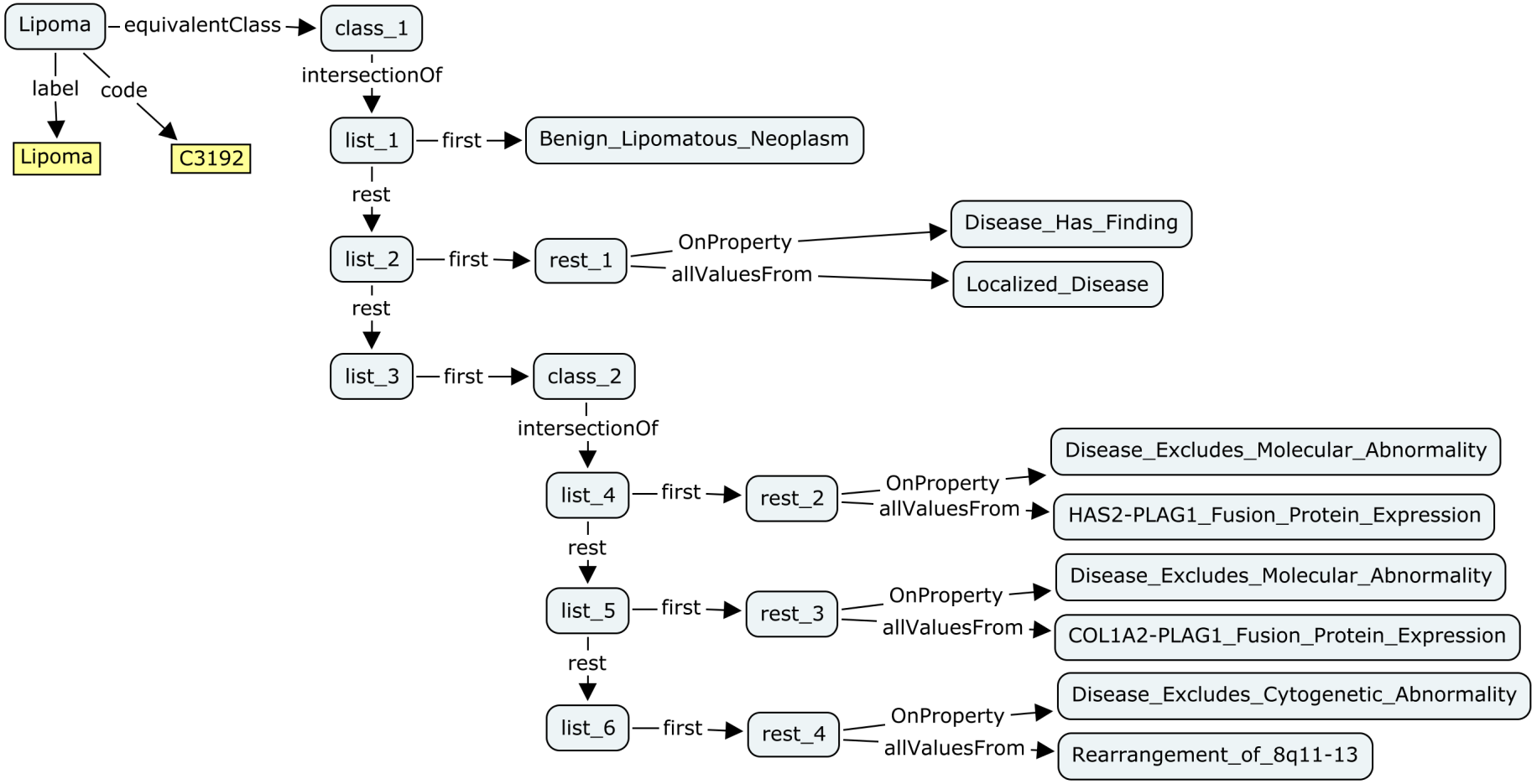
Subject	Predicate	Object
Lipoma	label	"Lipoma"
Lipoma	code	"C3192"
Lipoma	Disease_Excludes_Cytogenetic_Abnormality	Rearrangement_of_8q11-13
Lipoma	Disease_Excludes_Molecular_Abnormality	COL1A2-PLAG1_Fusion_Protein_Expression
Lipoma	Disease_Excludes_Molecular_Abnormality	HAS2-PLAG1_Fusion_Protein_Expression
Lipoma	Disease_Has_Finding	Localized_Disease

Lipoma ?predicate ?object .

Actual Query Results

Subject	Predicate	Object
Lipoma	label	“Lipoma”
Lipoma	code	“C3192”

Actual OWL Graph



Paths in SparQL

```
SELECT ?r1 ?r4
FROM <htp://example.com/ex>
WHERE
{
  ?r1 part ?r2 .
  ?r2 part ?r3 .
  ?r3 contains ?r4 .
}
```

Subject	Predicate	Object
A	part	B
B	part	C
C	part	D
D	contains	E

r1	r4
B	E

```
?r1 part/part/contains ?r4
```

Limitations of Paths in SparQL

```
SELECT ?r1 ?r4
FROM <http://example.com/ex>
WHERE
{
  ?r1  part  ?r2 .
  ?r2  part  ?r3 .
  ?r3  contains  ?r4 .
}
```

Subject	Predicate	Object
A	part	B
B	part	C
C	part	D
D	contains	E

What if we want to find the resources that are connected via any number of part properties followed by a single contains property?

Gleen

Extending SparQL to regular paths

Gleen

- Gleen adds “regular paths” to SparQL
  - Supports a regular expression type syntax for specifying matching path patterns
  - Implemented in ARQ (Jena SparQL processor)
  - Uses standard extension mechanism - Property Function

Gleen OnPath Property Function

- Property Function extension mechanism
  - Supports calls to external Java functions
  - Enables custom triple matching
- OnPath function call
  - subject gleen:OnPath (“pathExpr” object) .

Sample Gleen Expressions

Symbols:

`?` : zero or one

`*` : zero or more

`+` : one or more

`/` : concatenation

`|` : alternation

`[]` : property delimiters

`()` : grouping operators

Lipoma gleen:OnPath ("[subClassOf]+" ?super) .

- Bind to the variable ?super all superclasses, recursively, of Lipoma

list_1 gleen:OnPath ("[rest]*/[first]" ?list_elem) .

- Bind to ?list_elem the elements of list_1

Gleen Lipoma Query

```
SELECT ?prop ?val
FROM <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl>
WHERE{
  nci:Lipoma  gleen:OnPath ( □  ?restriction ) .
   ?restriction  owl:onProperty  ?prop .
   ?restriction  owl:allValuesFrom ?val .
}
```

"([owl:equivalentClass]?/[owl:intersectionOf]/[rdf:rest]*/[rdf:first])+"

Gleen Lipoma Query Results

prop	val
nci:Disease_Excludes_Cytogenetic_Abnormality	nci:Rearrangement_of_8q11-13
nci:Disease_Excludes_Molecular_Abnormality	nci:COL1A2-PLAG1_Fusion_Protein_Expression
nci:Disease_Excludes_Molecular_Abnormality	nci:HAS2-PLAG1_Fusion_Protein_Expression
nci:Disease_Has_Finding	nci:Localized_Disease
nci:Disease_Excludes_Normal_Cell_Origin	nci:Neuron_and_Supporting_Cell_of_the_Nervous_System
nci:Disease_Has_Normal_Tissue_Origin	nci:Adipose_Tissue
nci:Disease_Has_Normal_Tissue_Origin	nci:Connective_and_Soft_Tissue
nci:Disease_Has_Abnormal_Cell	nci:Neoplastic_Connective_and_Soft_Tissue_Cell
nci:Disease_Has_Normal_Cell_Origin	nci:Lipocyte
nci:Disease_Has_Normal_Cell_Origin	nci:Connective_and_Soft_Tissue_Cell
nci:Disease_Has_Finding	nci:Indolent_Clinical_Course
nci:Disease_Has_Abnormal_Cell	nci:Neoplastic_Lipocyte

Creating a View

Lipoma ?predicate ?object

Lipoma View Query

~~SELECT ?prop ?val~~

```
CONSTRUCT { nci:Lipoma ?prop ?val . }
FROM <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl>
WHERE{
  nci:Lipoma  gleen:OnPath (
    "([owl:equivalentClass]?/[owl:intersectionOf]/[rdf:rest]*/[rdf:first])+"  ?restriction ) .
  ?restriction  owl:onProperty  ?prop .
  ?restriction  owl:allValuesFrom  ?val .
}
```

Simplified Lipoma View

```
<rdf:RDF
    xmlns:nci="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#"
    xmlns:gleen="java:edu.washington.sig.gleen."
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#" >
  <rdf:Description rdf:about="#Lipoma">
    <nci:Disease_Has_Normal_Cell_Origin rdf:resource="#Connective_and_Soft_Tissue_Cell"/>
    <nci:Disease_Has_Normal_Tissue_Origin rdf:resource="#Adipose_Tissue"/>
    <nci:Disease_Excludes_Molecular_Abnormality rdf:resource="#HAS2-PLAG1_Fusion_Protein_Expression"/>
    <nci:Disease_Has_Finding rdf:resource="#Indolent_Clinical_Course"/>
    <nci:Disease_Has_Normal_Tissue_Origin rdf:resource="#Connective_and_Soft_Tissue"/>
    <nci:Disease_Excludes_Normal_Cell_Origin rdf:resource="#Neuron_and_Supporting_Cell_of_the_Nervous_
    <nci:Disease_Excludes_Molecular_Abnormality rdf:resource="#COL1A2-PLAG1_Fusion_Protein_Expression"
    <nci:Disease_Has_Finding rdf:resource="#Localized_Disease"/>
    <nci:Disease_Has_Normal_Cell_Origin rdf:resource="#Lipocyte"/>
    <nci:Disease_Has_Abnormal_Cell rdf:resource="#Neoplastic_Connective_and_Soft_Tissue_Cell"/>
    <nci:Disease_Has_Abnormal_Cell rdf:resource="#Neoplastic_Lipocyte"/>
    <nci:Disease_Excludes_Cytogenetic_Abnormality rdf:resource="#Rearrangement_of_8q11-13"/>
  </rdf:Description>
</rdf:RDF>
```

Simplified Lipoma View

```
<rdf:RDF
    xmlns:nci="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#"
    xmlns:gleen="java:edu.washington.sig.gleen."
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#" >
  <rdf:Description rdf:about="#Lipoma">
    <nci:Disease_Has_Normal_Cell_Origin rdf:resource="#Connective_and_Soft_Tissue_Cell"/>
    <nci:Disease_Has_Normal_Tissue_Origin rdf:resource="#Adipose_Tissue"/>
    <nci:Disease_Excludes_Molecular_Abnormality rdf:resource="#HAS2-PLAG1_Fusion_Protein_Expression"/>
    <nci:Disease_Has_Finding rdf:resource="#Indolent_Clinical_Course"/>
    <nci:Disease_Has_Normal_Tissue_Origin rdf:resource="#Connective_and_Soft_Tissue"/>
    <nci:Disease_Excludes_Normal_Cell_Origin rdf:resource="#Neuron_and_Supporting_Cell_of_the_Nervous_
    <nci:Disease_Excludes_Molecular_Abnormality rdf:resource="#COL1A2-PLAG1_Fusion_Protein_Expression"
    <nci:Disease_Has_Finding rdf:resource="#Localized_Disease"/>
    <nci:Disease_Has_Normal_Cell_Origin rdf:resource="#Lipocyte"/>
    <nci:Disease_Has_Abnormal_Cell rdf:resource="#Neoplastic_Connective_and_Soft_Tissue_Cell"/>
    <nci:Disease_Has_Abnormal_Cell rdf:resource="#Neoplastic_Lipocyte"/>
    <nci:Disease_Excludes_Cytogenetic_Abnormality rdf:resource="#Rearrangement_of_8q11-13"/>
  </rdf:Description>
</rdf:RDF>
```

Query on View

```
SELECT ?predicate ?object
FROM <http://.../View.rdf>
WHERE
{
  nci:Lipoma  ?predicate  ?object .
}
```

predicate	object
nci:Disease_Excludes_Cytogenetic_Abnormality	nci:Rearrangement_of_8q11-13
nci:Disease_Excludes_Molecular_Abnormality	nci:COL1A2-PLAG1_Fusion_Protein_Expression
nci:Disease_Excludes_Molecular_Abnormality	nci:HAS2-PLAG1_Fusion_Protein_Expression
nci:Disease_Has_Finding	nci:Localized_Disease
nci:Disease_Excludes_Normal_Cell_Origin	nci:Neuron_and_Supporting_Cell_of_the_Nervous_System
nci:Disease_Has_Normal_Tissue_Origin	nci:Adipose_Tissue
nci:Disease_Has_Normal_Tissue_Origin	nci:Connective_and_Soft_Tissue
nci:Disease_Has_Abnormal_Cell	nci:Neoplastic_Connective_and_Soft_Tissue_Cell
nci:Disease_Has_Normal_Cell_Origin	nci:Lipocyte
nci:Disease_Has_Normal_Cell_Origin	nci:Connective_and_Soft_Tissue_Cell
nci:Disease_Has_Finding	nci:Indolent_Clinical_Course
nci:Disease_Has_Abnormal_Cell	nci:Neoplastic_Lipocyte

Related Work

- Alkhateeb F, Baget J-F, Euzenat J. Extending SPARQL with regular expression patterns. Institut National de Recherche en Informatique et Automatique (INRIA), Tech Rep 6191. 2007.
- Kochut K, Janik M. SPARQLeR: Extended Sparql for Semantic Association Discovery. European Semantic Web Conference (ESWC). 2007:145-59.

Conclusion

- Query-based approach supports generation of simplified views of OWL ontologies
  - More intuitive for readers
  - Easier to query
- Gleen extends SparQL
  - Regular path patterns required
  - Gleen plugin provides processing for regular paths
- Future Work
  - Identify candidate patterns in OWL
  - Improve view generation efficiency

Acknowledgements

- Grant support:"Realizing the potential of reference ontologies for the semantic web" NIH grant HL087706
- The Jena and ARQ teams
  - Particular thanks to Andy Seaborne of HP Labs
  - ARQ now offers direct support for path expressions

http://sig.biostr.washington.edu/projects/ontviews/gleen