

The CALBC RDF Triple store: retrieval over large literature content

Samuel Croset, Christoph Grabmüller, Chen Li, Silverstras Kavaliauskas, Dietrich Rebholz-Schuhmann

croset@ebi.ac.uk

SamuelCroset.com



EMBL-EBI



Outline

- Motivation
- Integrating multiple resources
 - CALBC Corpus
 - LexEBI
 - Public databases
- Querying the Triple Store

Outline

- Motivation
- Integrating multiple resources
 - CALBC Corpus
 - LexEBI
 - Public databases
- Querying the Triple Store

Why representing scientific literature in RDF?

- Scientific literature:
 - Primary data resource reporting novel scientific findings
- Text-mining:
 - Biological entities recognition
 - Population of biomedical databases through curators
- RDF representation:
 - Standardization of the content extracted
 - **Exploitation of the literature in the Semantic Web**

Outline

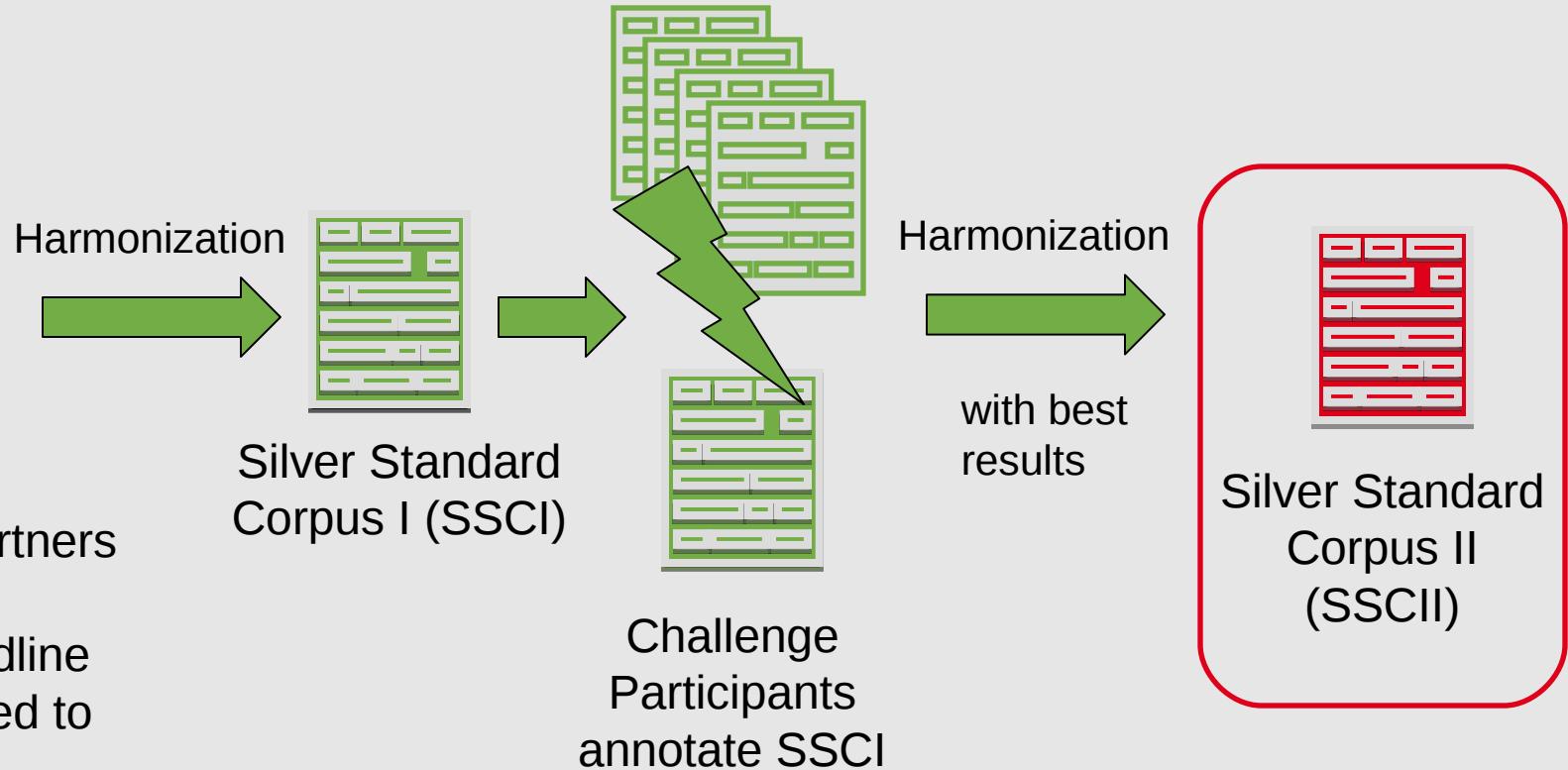
- Motivation
- Integrating multiple resources
 - CALBC Corpus
 - LexEBI
 - Public databases
- Querying the Triple Store

CALBC Corpus

- Collaborative Annotation of a Large Biomedical Corpus

• doi:10.1338/nph.2011.0001
• Proceedings of the National
Institutes of Health
• 4 Project partners

• 150'000 Medline
abstract related to
Immunology
annotated



CALBC Corpus

- Advantages of the CALBC Corpus:
- Large-scale corpus
- 4 semantic types: Gene-Protein, Diseases, Chemicals and Species
- Generated in a purely automatic way
- Highly reproducible
- <http://www.calbc.eu/>

CALBC in RDF

<http://www.ebi.ac.uk/Rebholz/core/calbc/sentenceid#1>

0605

Nature Proceedings : doi:10.1038/npre.2011.5383.2 : Posted 18 Jan 2011

calbc:hasSentence

<http://www.ncbi.nlm.nih.gov/pubmed/442>

92

dc:date

1980-06-16

calbc:isIn

dc:identifier

<urn:issn:0004-5772>

dc:creator

Seshadri, M
S

Varkey, K

dc:title

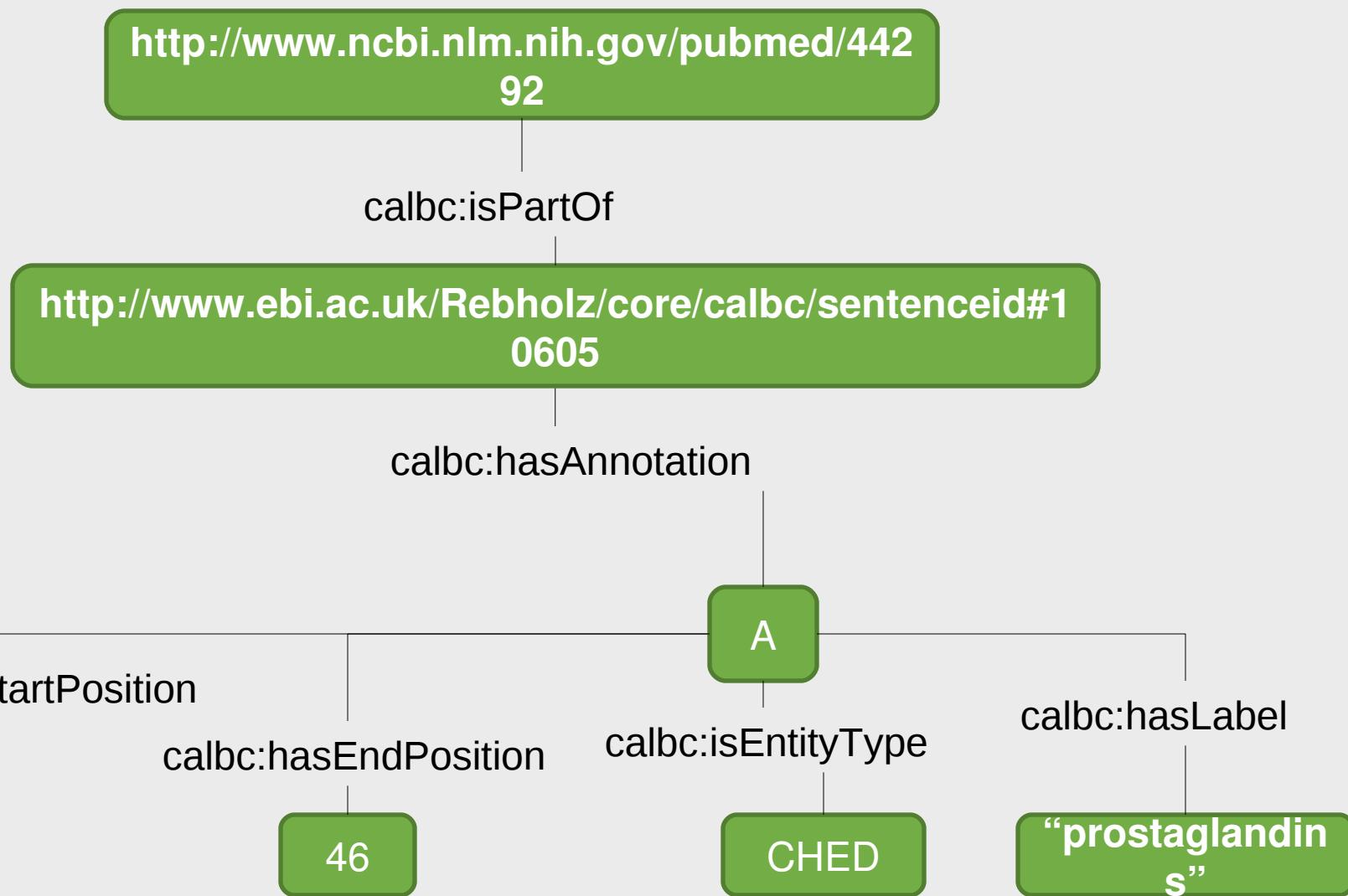
http://www.ebi.ac.uk/Rebholz/core/corpus_calbc

“Hepatitis B surface antigen (HBsAg) positive polyarteritis nodosa. A report of two cases and review of literature”

@prefix dc: <http://purl.org/dc/elements/1.1/>

@prefix calbc: <http://www.ebi.ac.uk/Rebholz/core/calbc/>

CALBC in RDF



@prefix calbc: <http://www.ebi.ac.uk/Rebholz/core/calbc/>

Outline

- Motivation
- Integrating multiple resources
 - CALBC Corpus
 - LexEBI
 - Public databases
- Querying the Triple Store

LexEBI

- BioThesaurus: Complete term repository for the biomedical domain
- LexEBI → XML
- Features:
 - Frequency count for the occurrence of the term in British National Corpus (BNC) or in MEDLINE → **Disambiguation**
 - Mapping to original resource (URI) → **Normalization**

LexEBI in RDF

16

http://purl.org/obo/owl/NCBITaxon_338969

:species

<http://purl.uniprot.org/uniprot/Q21W>

W8

:isIn

http://www.ebi.ac.uk/Rebholz/core/corpus_lexebi

:hasVariant

A

:FrequencyInMedline

30

:preferredTerm

True

A

:hasVariant

:isType

:isType

:SurfaceForm

“Ribosomal large
subunit
methyltransferase E”

:FrequencyInMedline

:SurfaceForm

“ftsj”

Orthographic

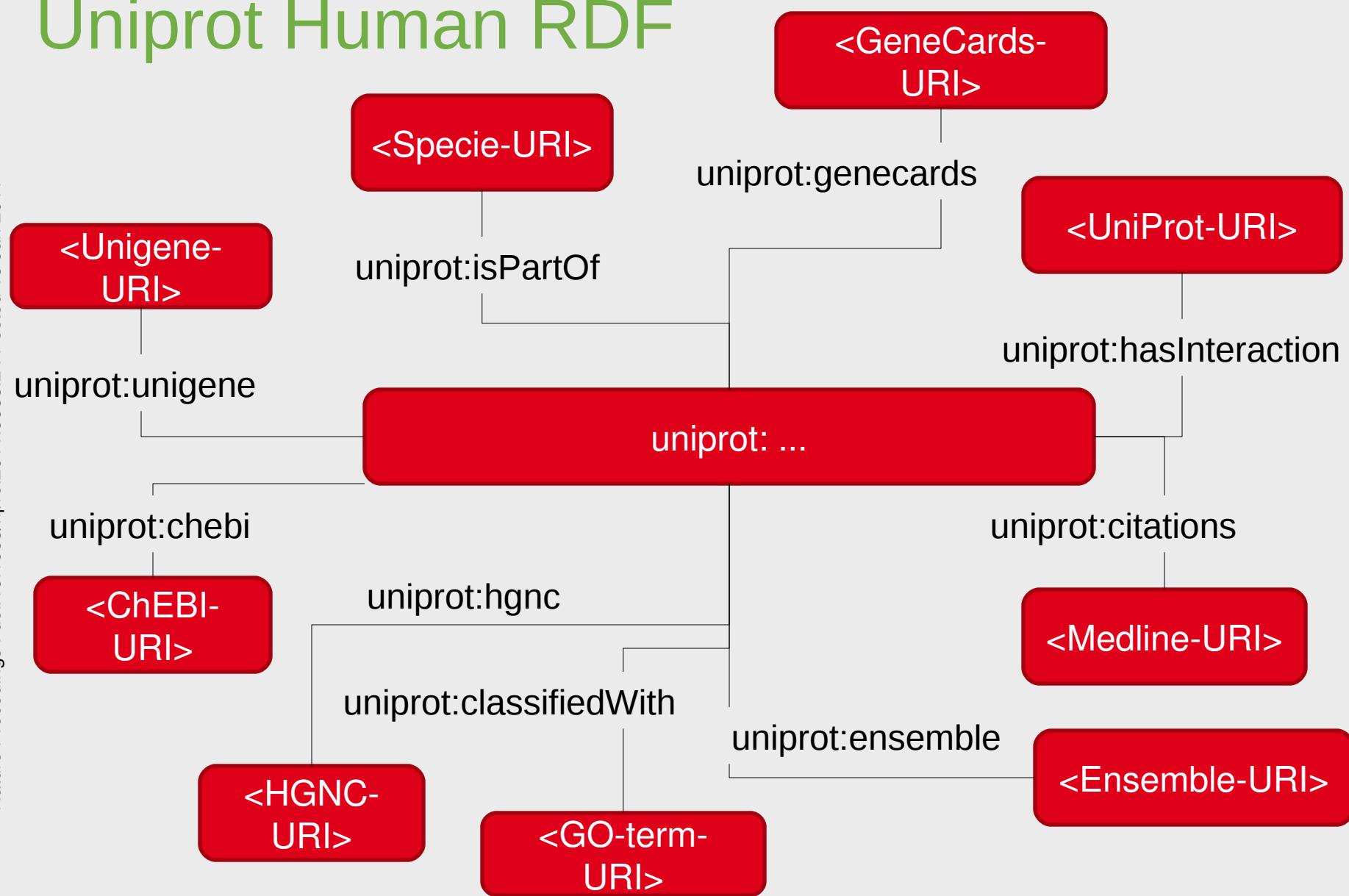


Outline

- Motivation
- Integrating multiple resources
 - CALBC Corpus
 - LexEBI
 - Public databases
- Querying the Triple Store

Uniprot Human RDF

Nature Precedings : doi:10.1038/npre.2011.5383.2 : Posted 18 Jan 2011



@prefix uniprot: <http://purl.uniprot.org/uniprot/>

Uniprot RDF

Nature Precedings : doi:10.1038/npre.2011.5383.2 : Posted 18 Jan 2011



@prefix uniprot: <http://purl.uniprot.org/uniprot/>

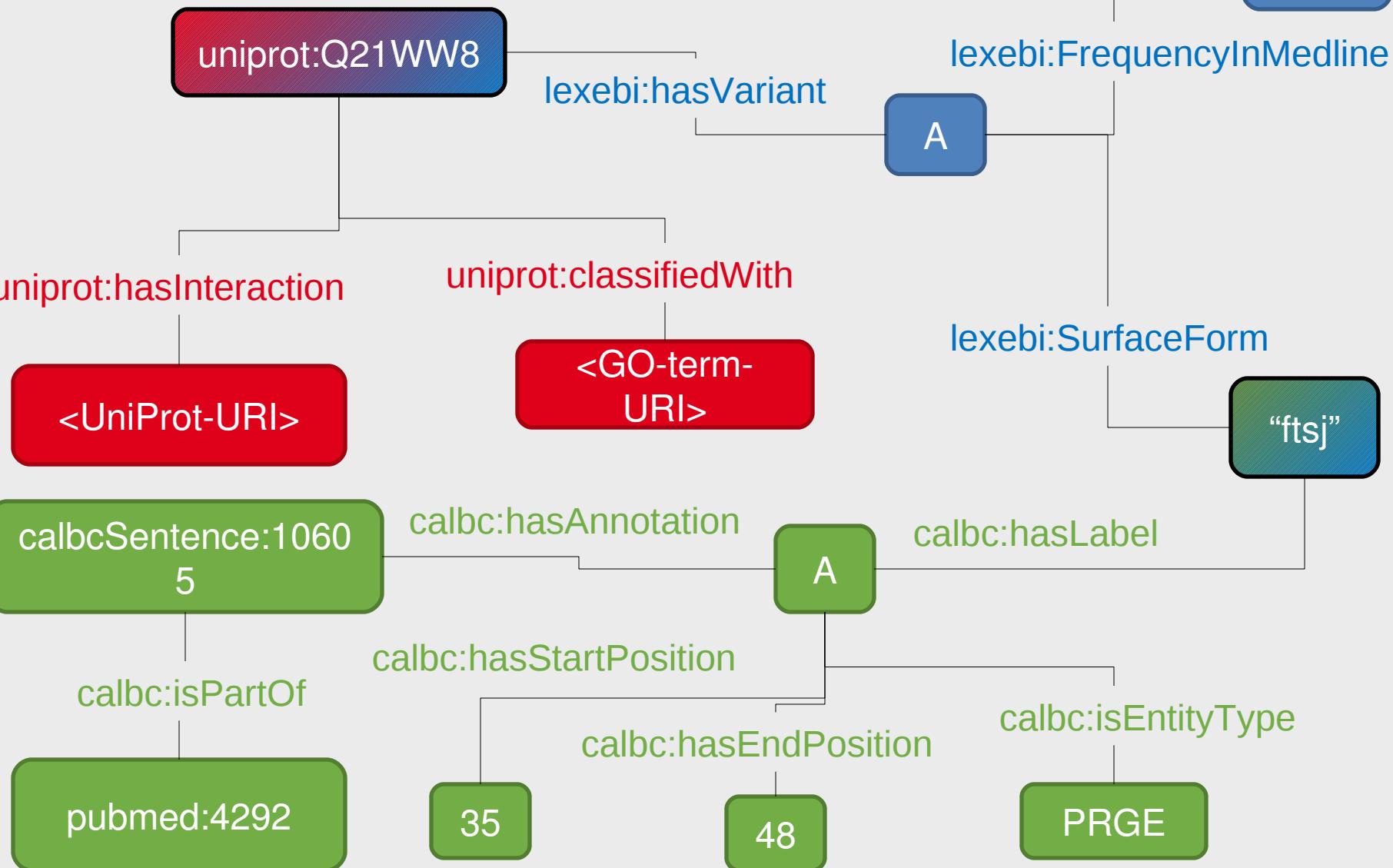
Outline

- Motivation
- Integrating multiple resources
 - CALBC Corpus
 - LexEBI
 - Public databases
- Querying the Triple Store

Querying the Triple Store

16

Nature Precedings · doi:10.1038/npre.2011.5383.2 : Posted 18 Jan 2011



Use cases

- Normalization of CALBC named entities
- Disambiguation of CALBC named entities
- Term collocation at the sentence level → e.g.
Evidence for Gene – Disease association
- Checking consistency of bioinformatics resources
from literature

Thank you for your
attention