Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

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

SWAT4LS

EMBL-EBI

10th December 2010, Berlin

EBI is an Outstation of the European Molecular Biology Laboratory.

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

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

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

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**

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

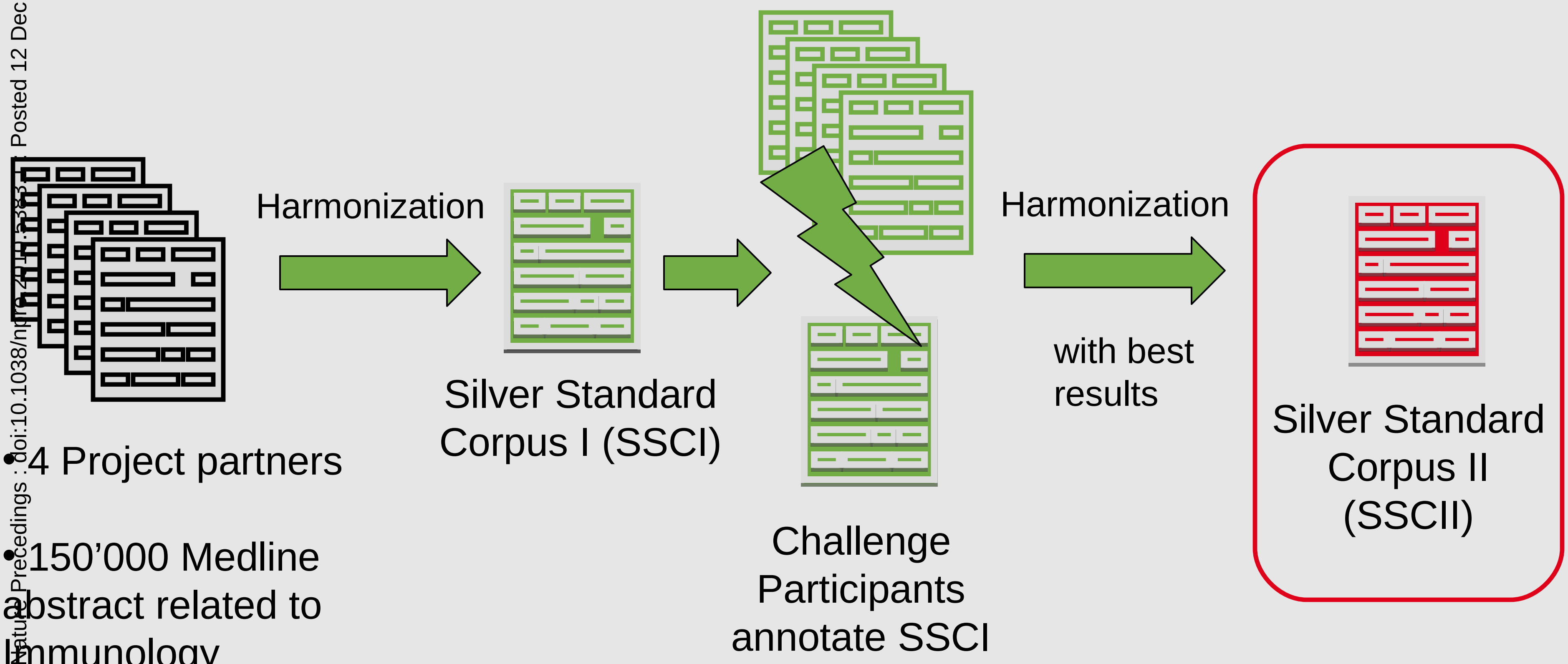
Outline

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

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

CALBC Corpus

- Collaborative Annotation of a Large Biomedical Corpus



Harmonization

Silver Standard Corpus I (SSCI)

Challenge Participants annotate SSCI

Harmonization

with best results

Silver Standard Corpus II (SSCII)

- 4 Project partners

- 150’000 Medline abstract related to Immunology annotated

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

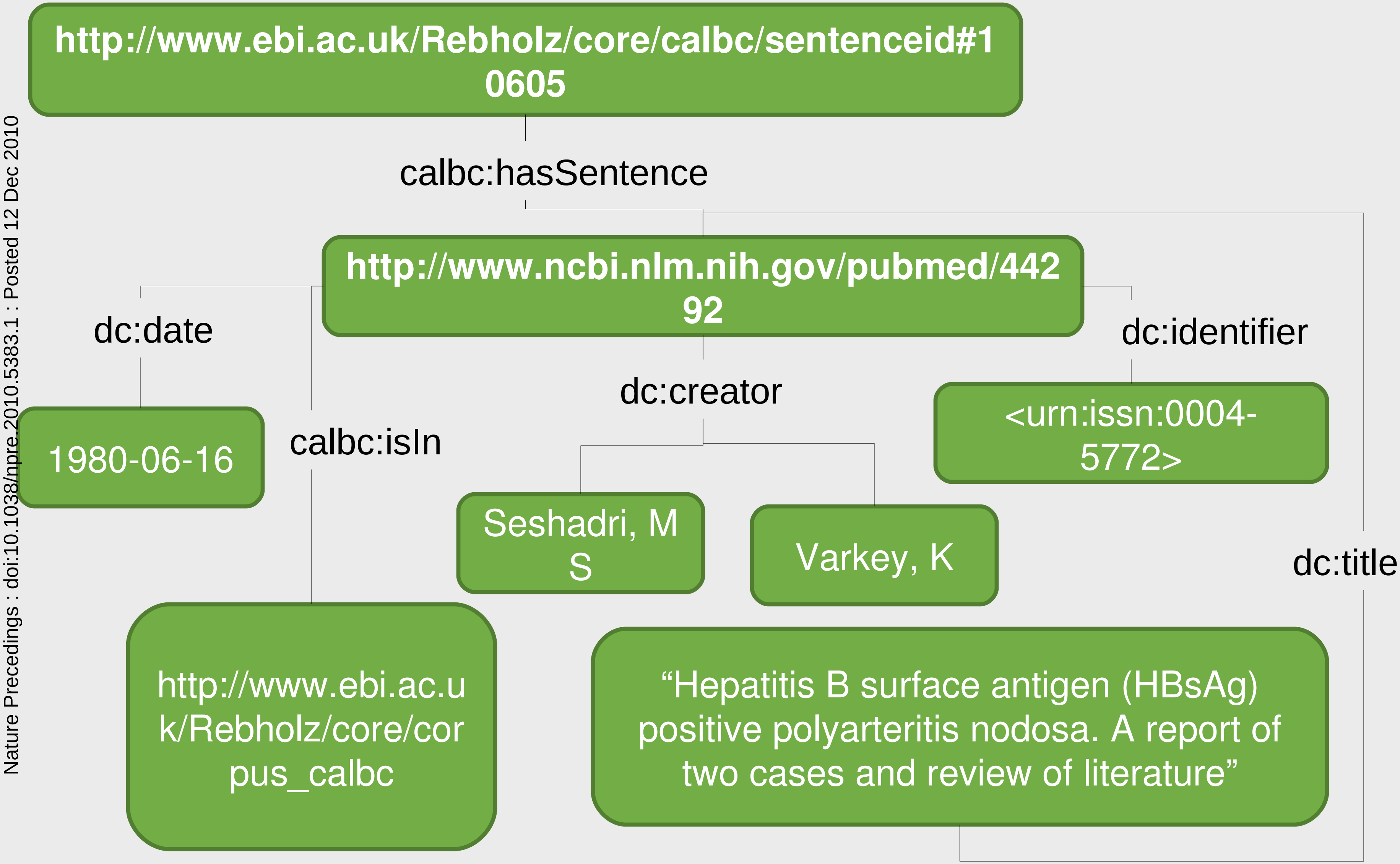
CALBC Corpus

- 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/

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

CALBC in RDF



http://www.ebi.ac.uk/Rebholz/core/calbc/sentenceid#10605

calbc:hasSentence

http://www.ncbi.nlm.nih.gov/pubmed/44292

dc:date

1980-06-16

calbc:isIn

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

dc:creator

Seshadri, M S

Varkey, K

dc:identifier

<urn:issn:0004-5772>

dc:title

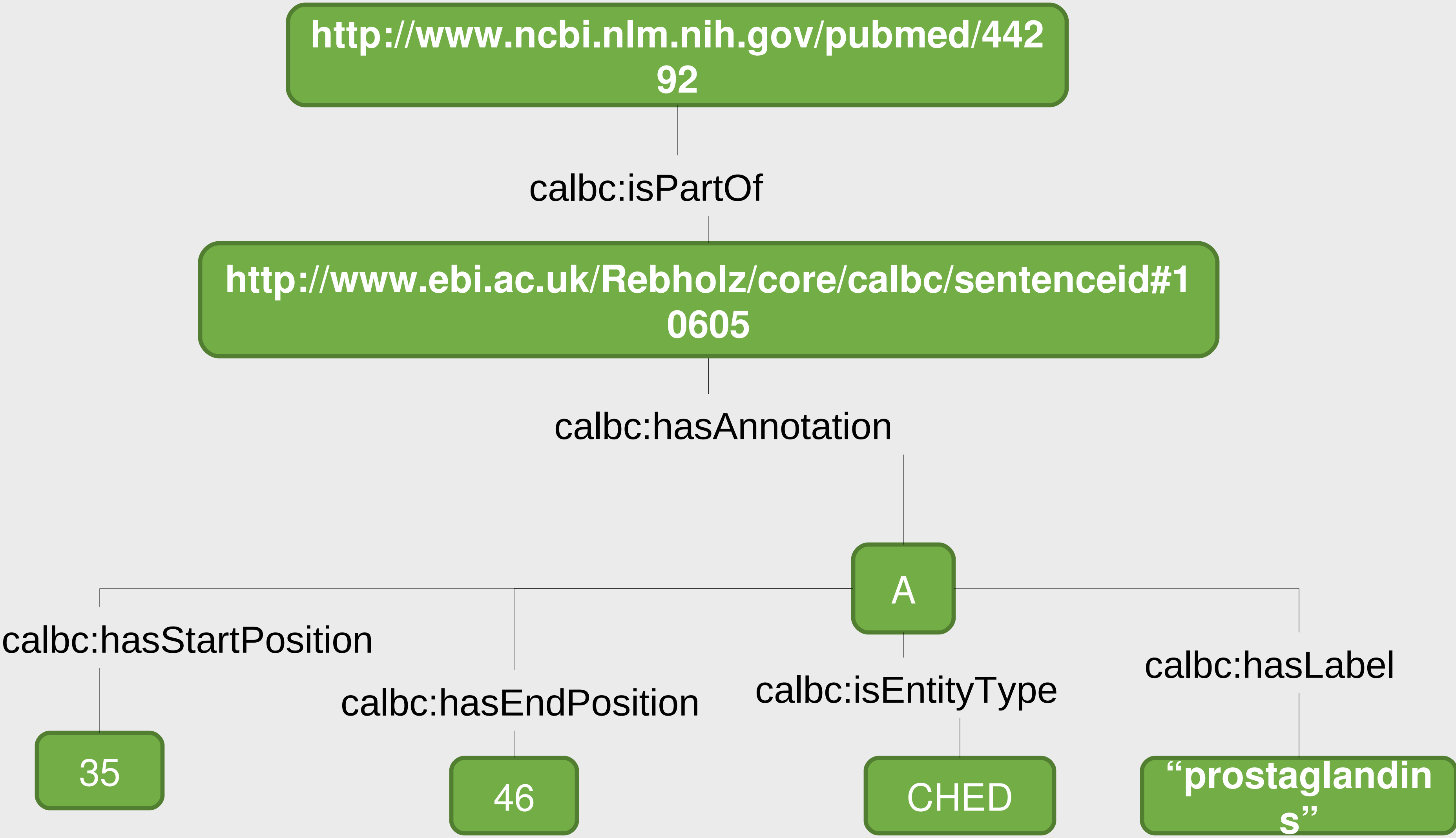
“Hepatitis B surface antigen (HBsAg) positive polyarteritis nodosa. A report of two cases and review of literature”

@prefix dc: http://purl.org/dc/elements/1.1/
@prefix calbc: http://www.ebi.ac.uk/Rebholz/core/calbc/

EMBL-EBI

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

CALBC in RDF



http://www.ncbi.nlm.nih.gov/pubmed/44292

calbc:isPartOf

http://www.ebi.ac.uk/Rebholz/core/calbc/sentenceid#10605

calbc:hasAnnotation

A

calbc:hasStartPosition

35

calbc:hasEndPosition

46

calbc:isEntityType

CHED

calbc:hasLabel

“prostaglandins”

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

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

Outline

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

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

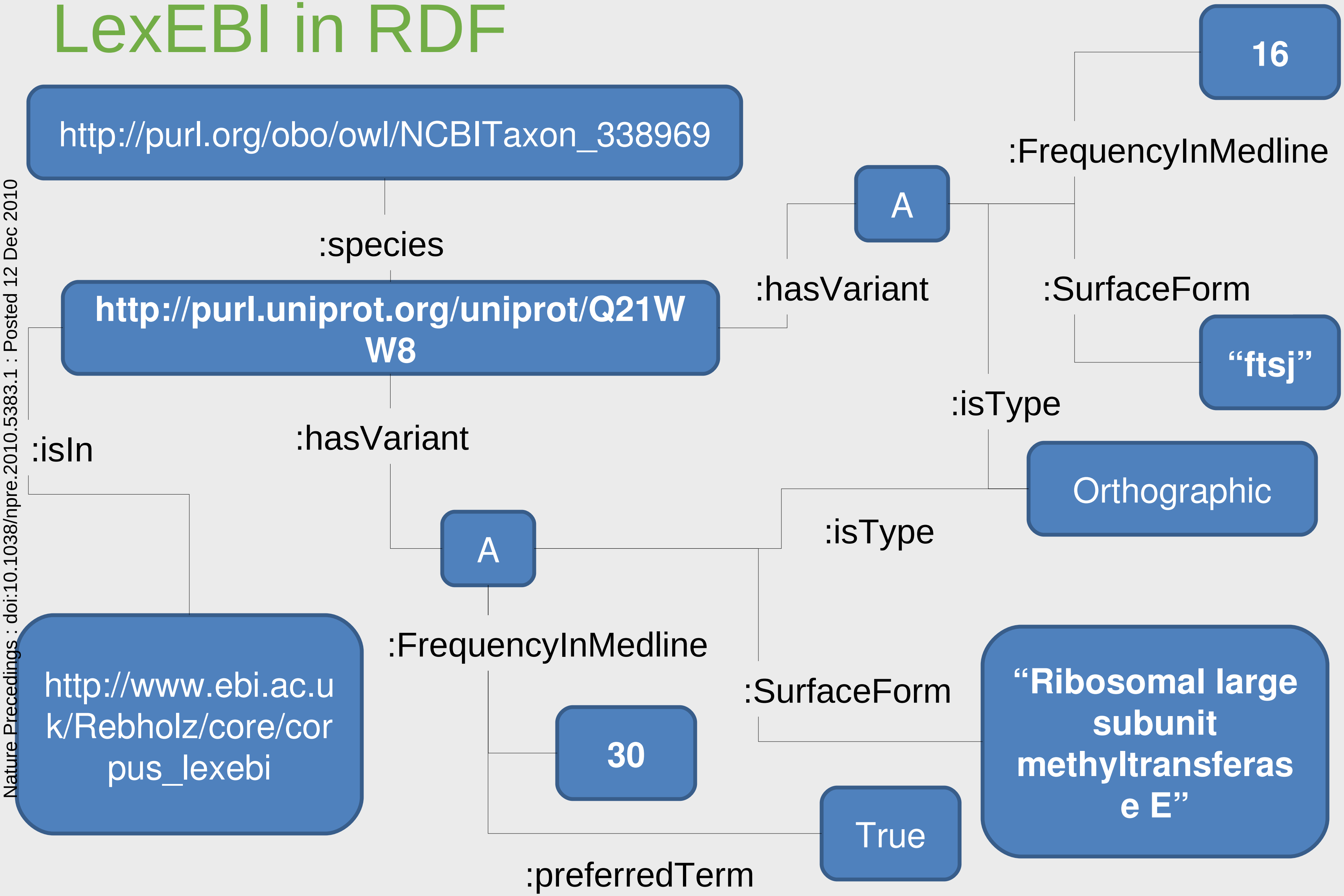
LexEBI

- BioThesaurus: Complete term repository for the biomedical domain
- LexEBI → XML

- <u>Features:</u>
  - Frequency count for the occurrence of the term in British National Corpus (BNC) or in MEDLINE → **Disambiguation**
  - Mapping to original resource (URI) → **Normalization**

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

LexEBI in RDF



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

:species

http://purl.uniprot.org/uniprot/Q21WW8

:isIn

:hasVariant

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

A

:FrequencyInMedline

30

:preferredTerm

True

:SurfaceForm

“Ribosomal large subunit methyltransferase E”

:isType

Orthographic

:hasVariant

A

:FrequencyInMedline

16

:SurfaceForm

“ftsj”

:isType

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

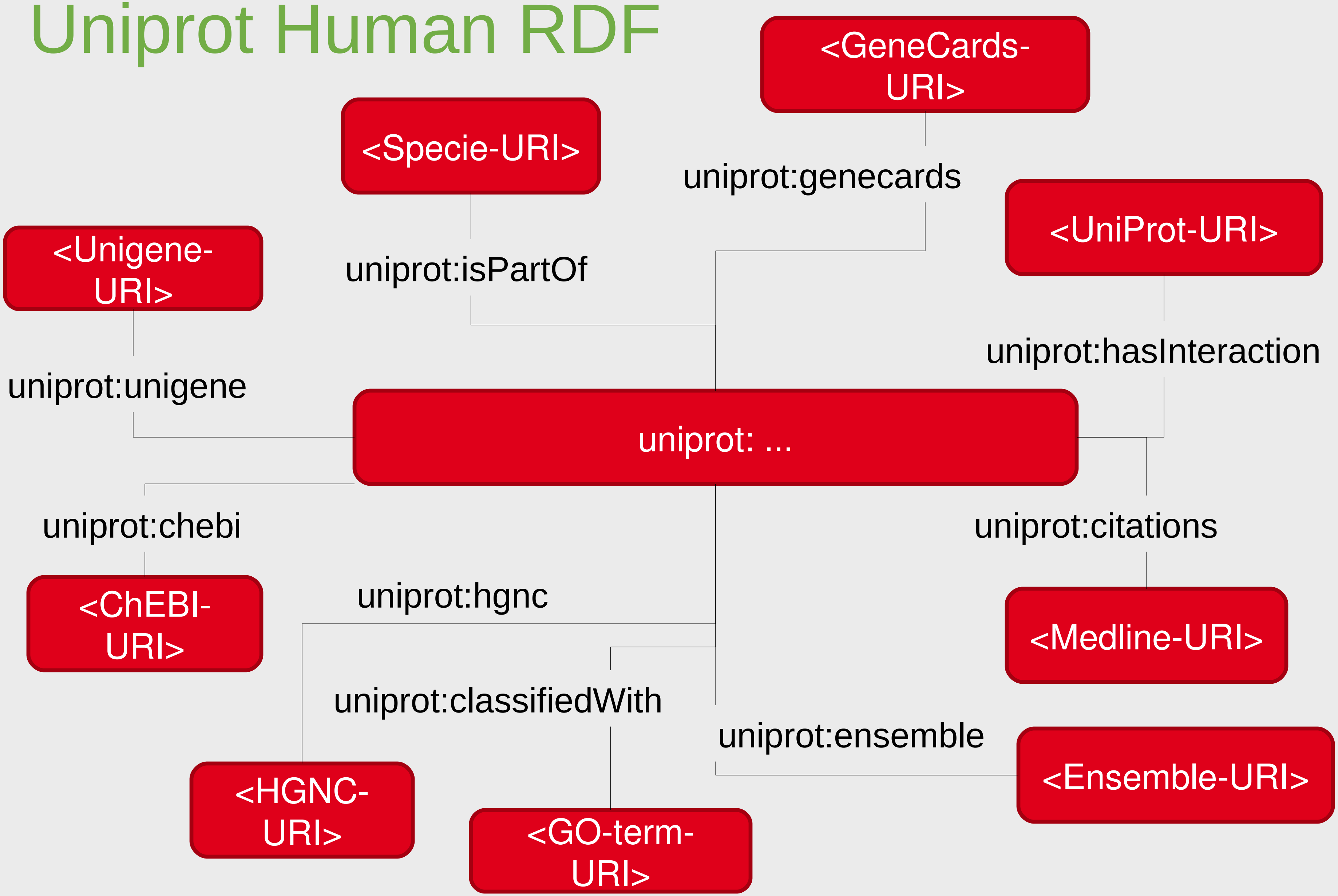
Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Outline

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

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Uniprot Human RDF

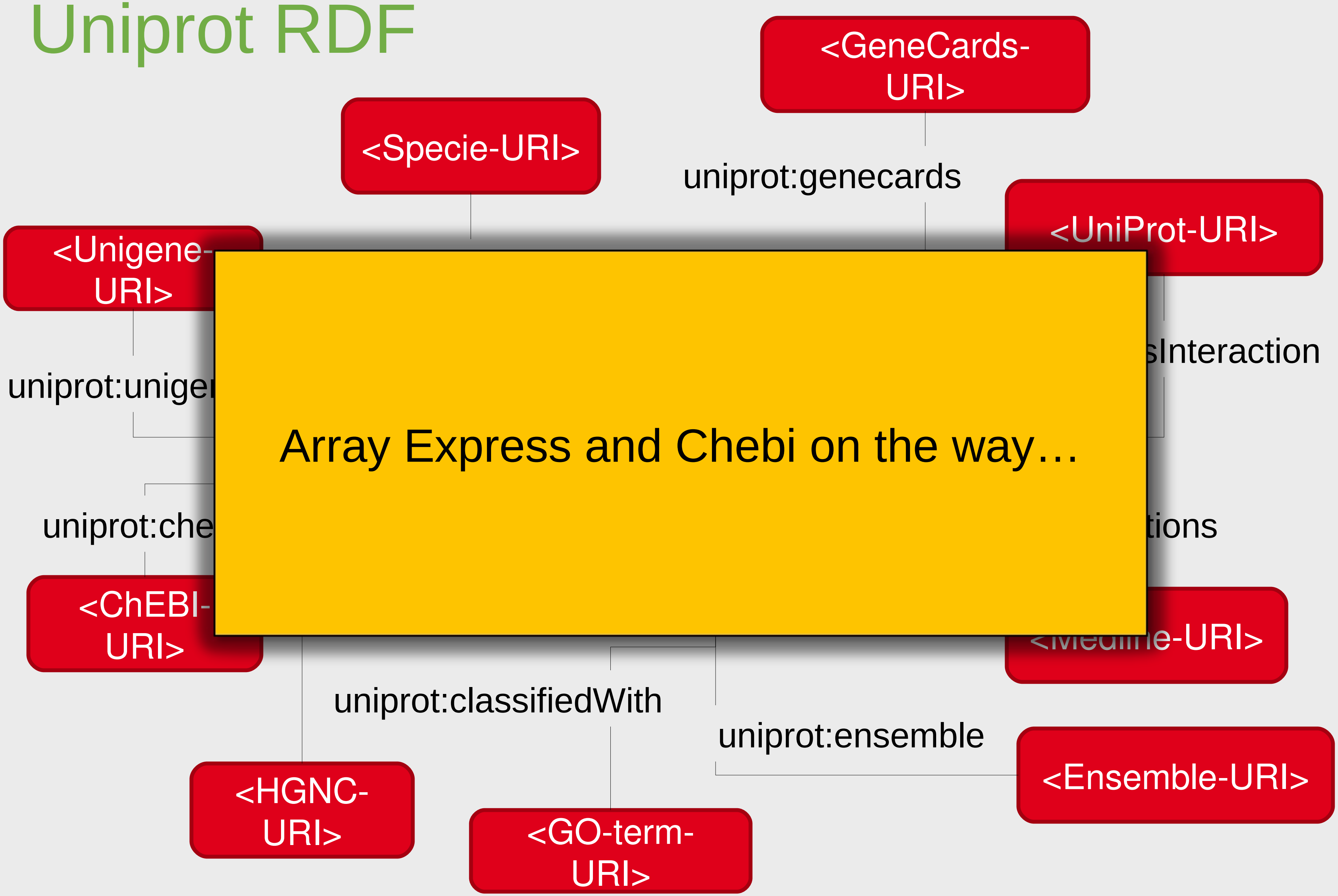


@prefix uniprot: http://purl.uniprot.org/uniprot/

EMBL-EBI

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Uniprot RDF



Array Express and Chebi on the way…

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

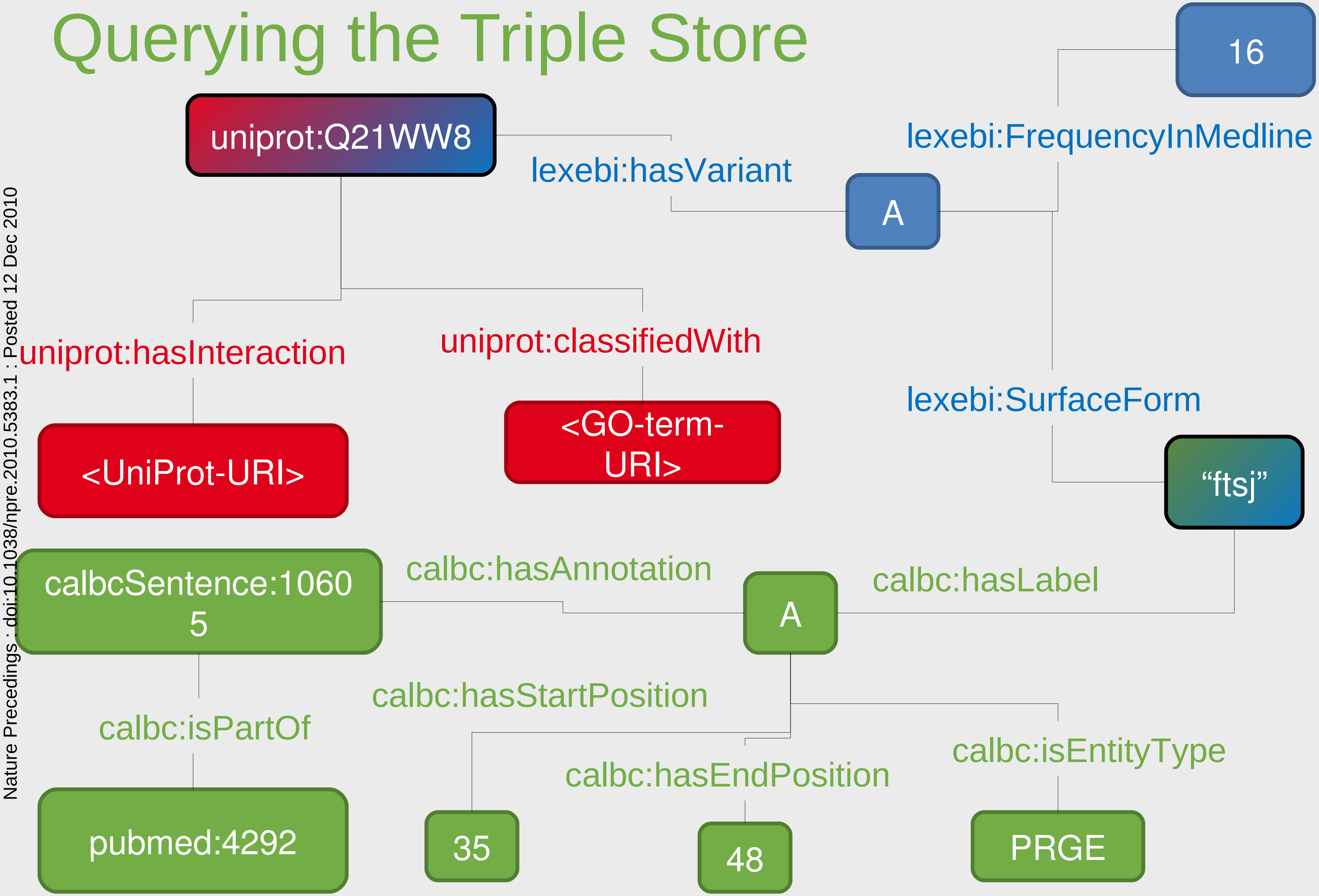
Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Outline

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

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Querying the Triple Store


uniprot:Q21WW8

lexebi:hasVariant

A

lexebi:FrequencyInMedline

16

uniprot:hasInteraction

<UniProt-URI>

uniprot:classifiedWith

<GO-term-URI>

lexebi:SurfaceForm

“ftsj”

calbcSentence:10605

calbc:hasAnnotation

A

calbc:hasLabel

calbc:isPartOf

pubmed:4292

calbc:hasStartPosition

35

calbc:hasEndPosition

48

calbc:isEntityType

PRGE

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

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

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Nature Precedings : doi:10.1038/npre.2010.5383.1 : Posted 12 Dec 2010

Thank you for your attention