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

10th December 2010, Berlin

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?

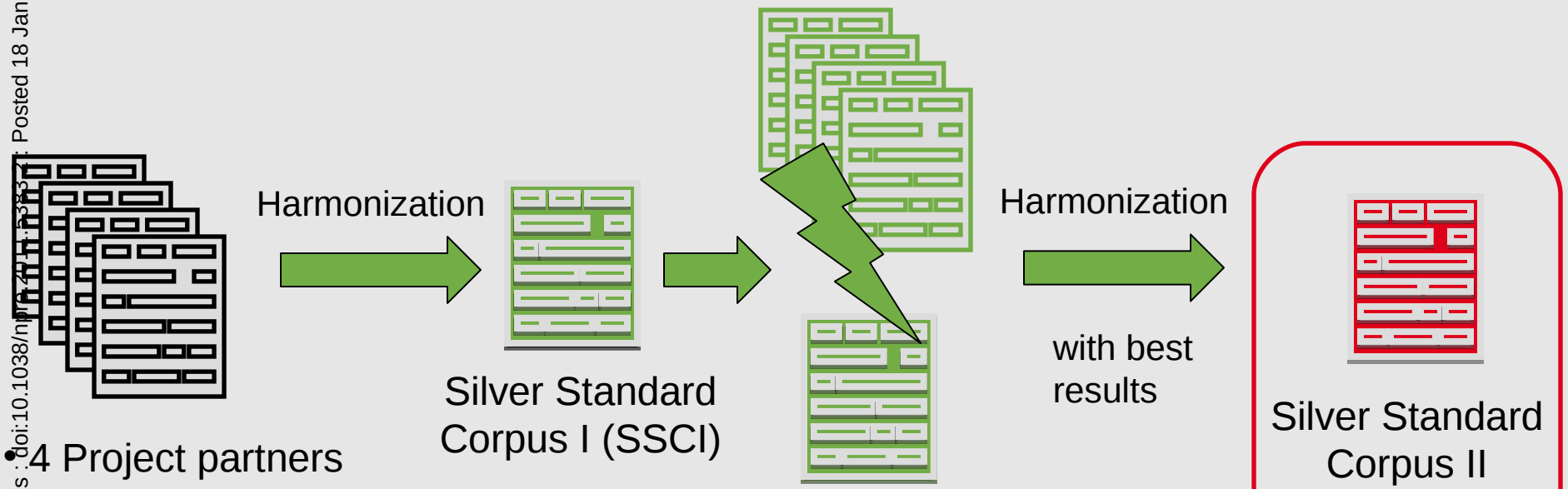
- Scientific literature:
  - Primary data resource reporting novel scientific findings
- Text-mining:
  - Biological entities recognition
  - Population of biomedical databases through curators
- RDF representation:
  - Standardization of the content extracted
  - **Exploitation of the literature in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



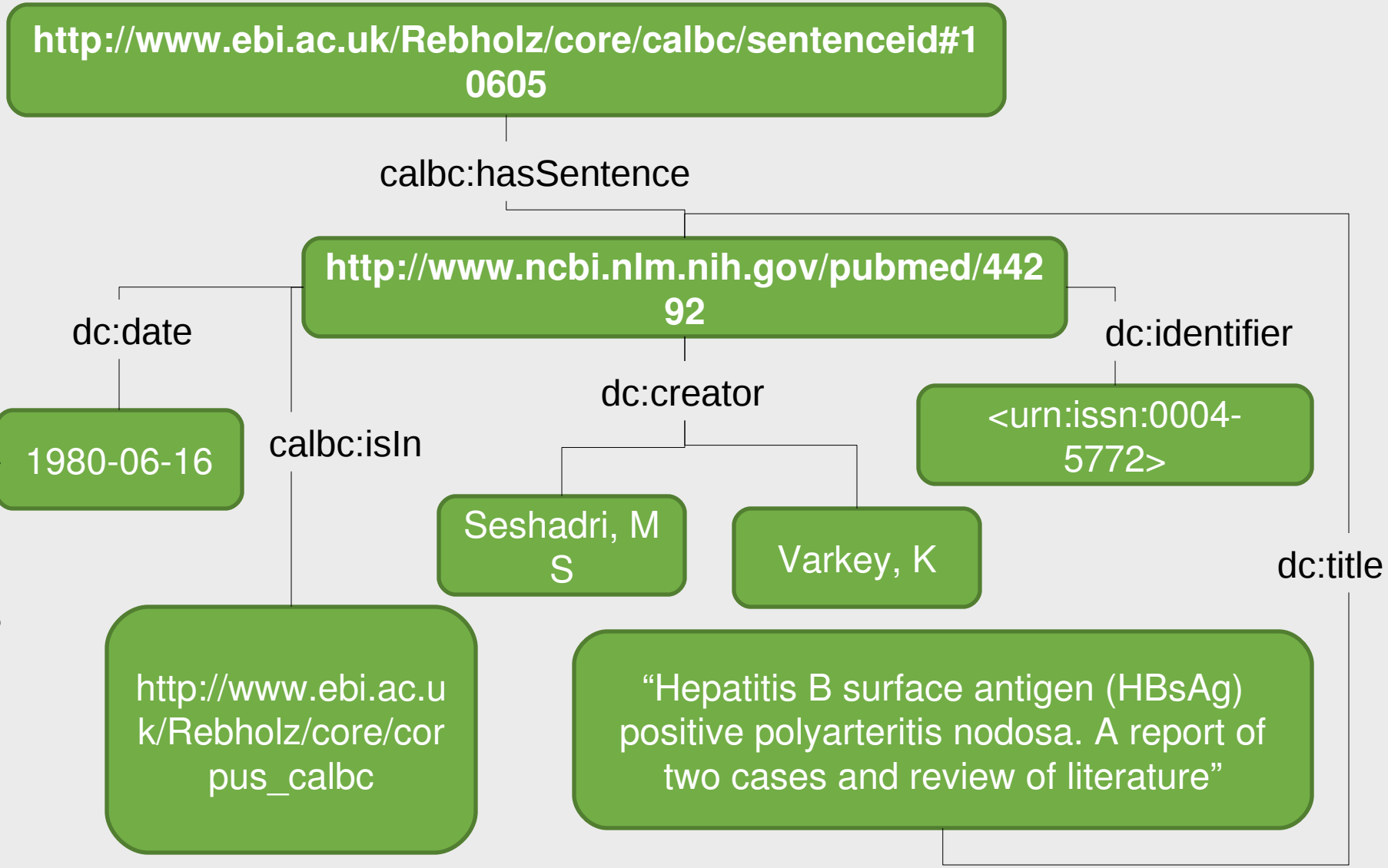
- 4 Project partners
- 150’000 Medline abstract related to Immunology annotated

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/

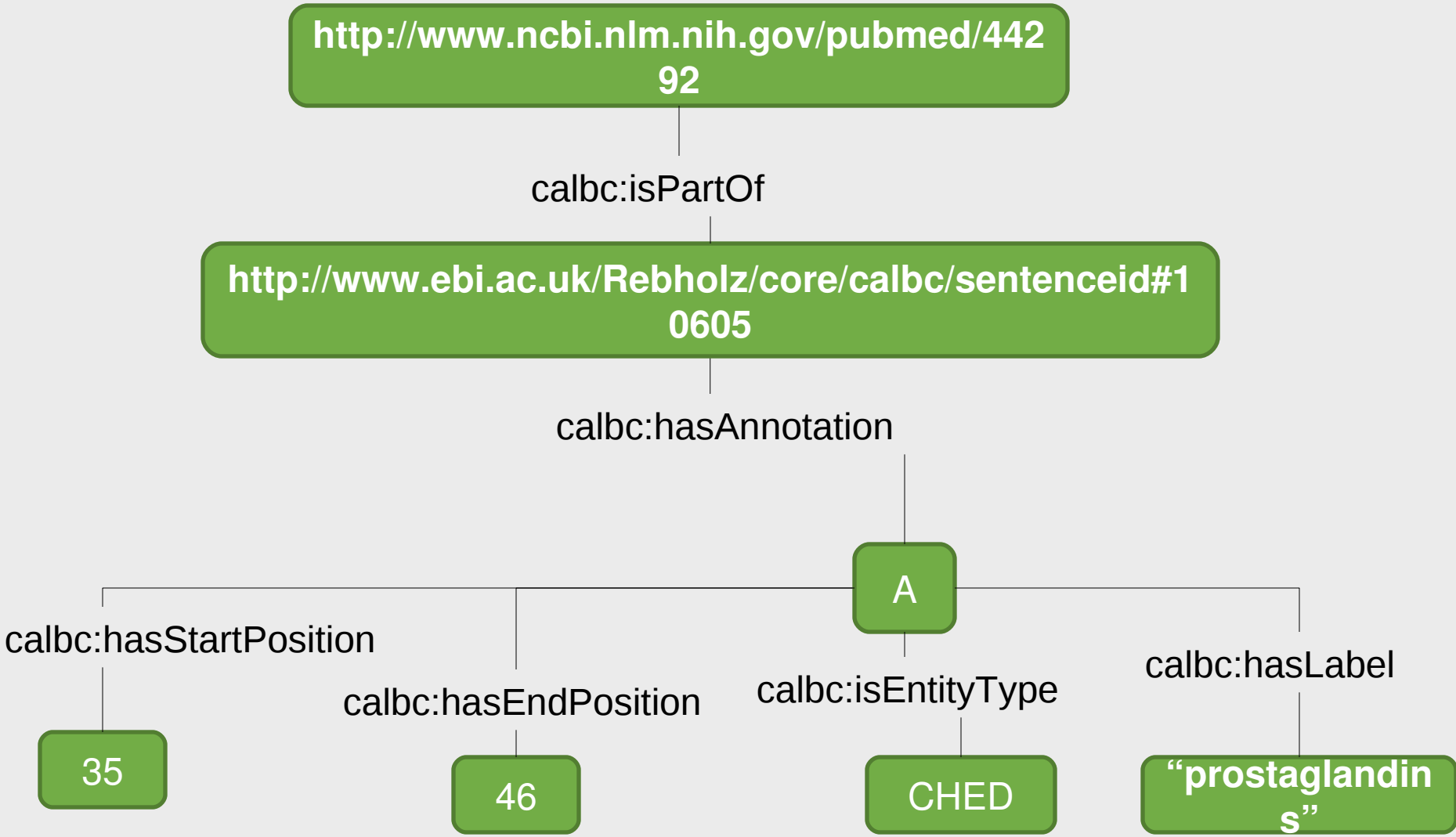
CALBC in RDF



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

EMBL-EBI

CALBC in RDF



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

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

EMBL-EBI

Outline

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

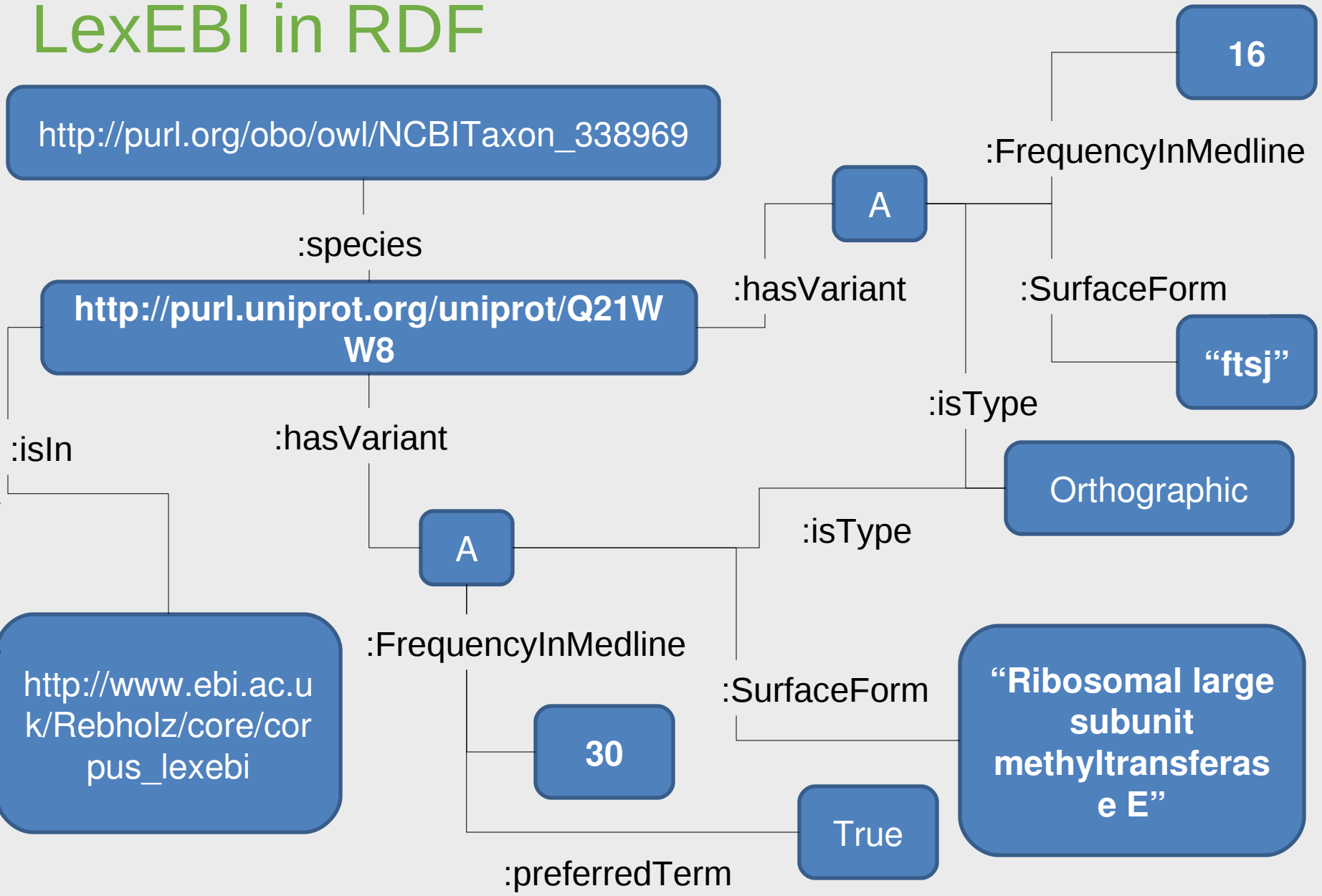
LexEBI

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

- <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**

LexEBI in RDF



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

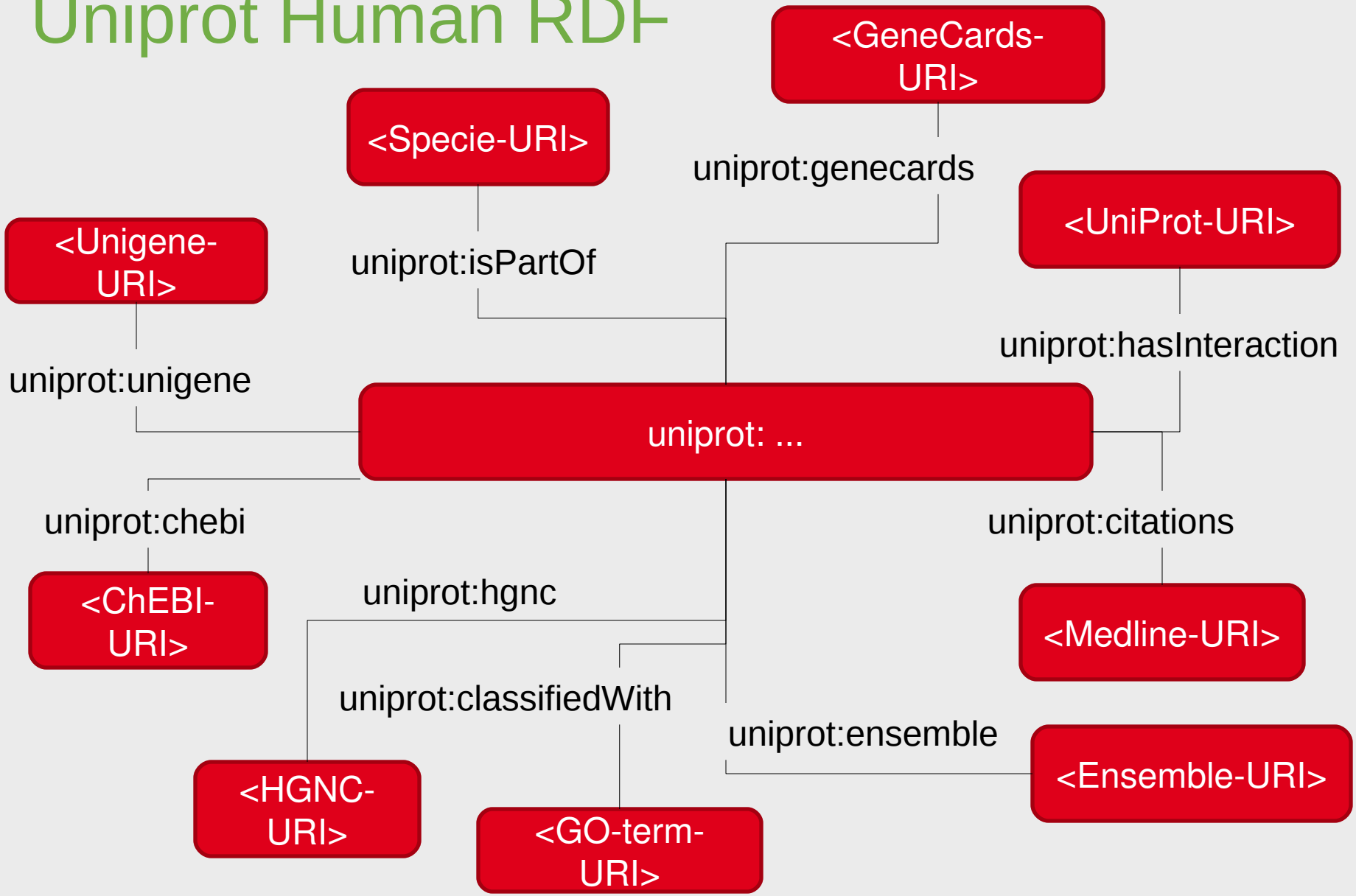
@prefix : <http://www.ebi.ac.uk/Rebholz/core/lexebi#>

EMBL-EBI

Outline

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

Uniprot Human RDF

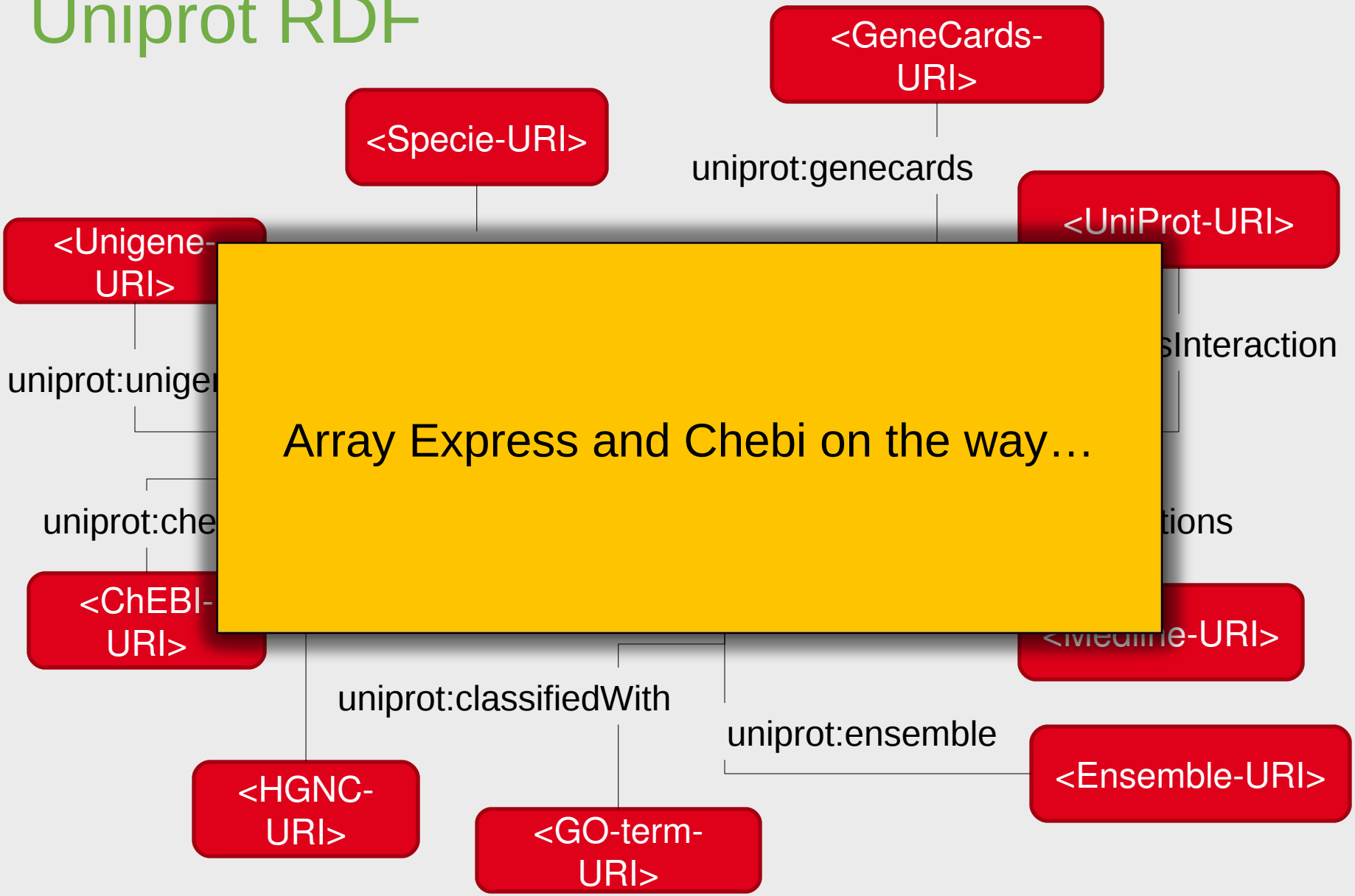


Nature Precedings : doi:10.1038/npre.2011.5383.2 : Posted 18 Jan 2011

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

EMBL-EBI

Uniprot RDF



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

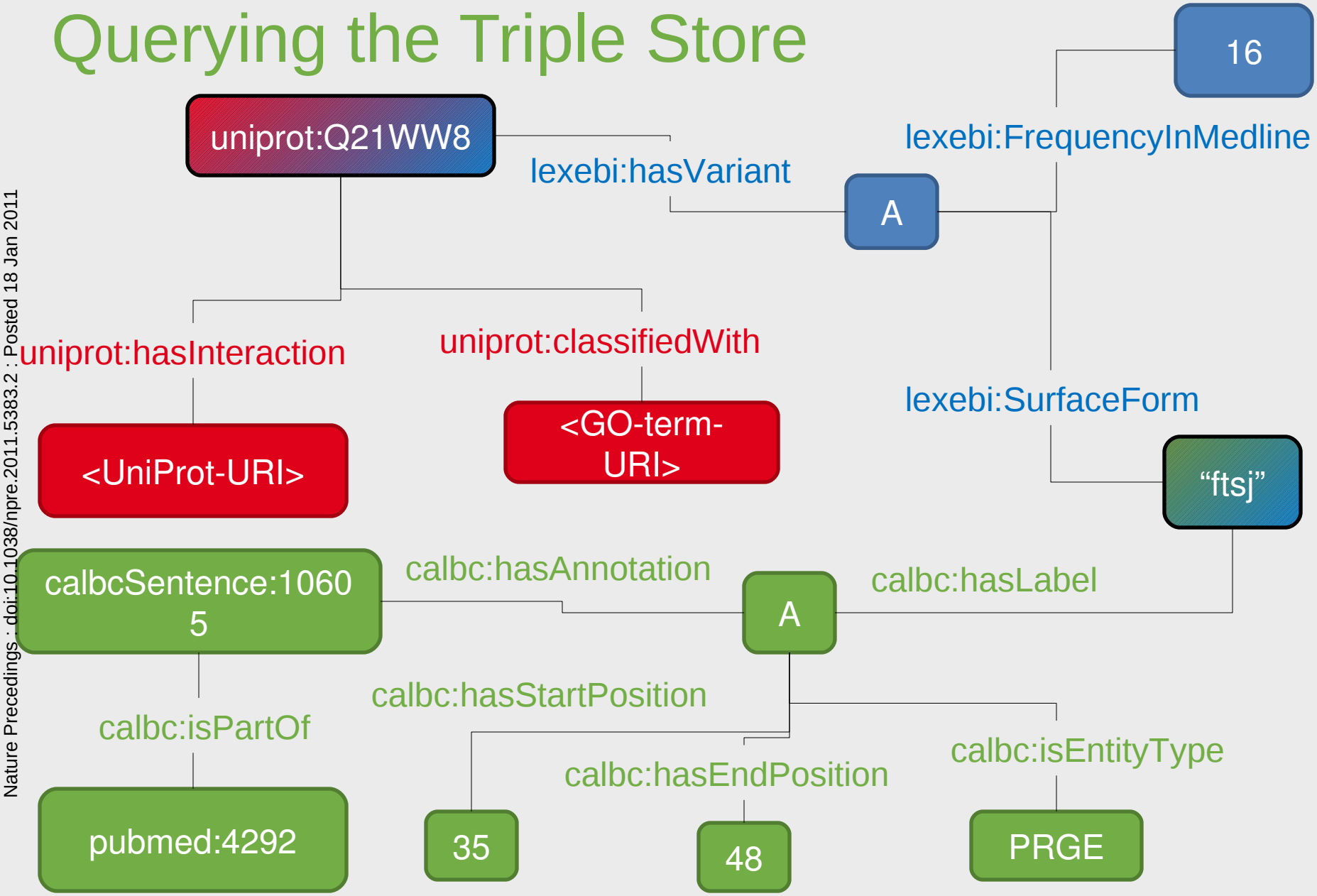
@prefix uniprot: http://purl.uniprot.org/uniprot/

EMBL-EBI

Outline

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

Querying the Triple Store



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

EMBL-EBI