UniDive

2nd General Meeting
University of Naples “L’Orientale
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Enhancing Interoperability for Under-Resourced Languages

A Case Study on Linking Lithuanian-English Data in the Cybersecurity Domain

Christian Chiarcos[1] Maxim Ionov[2] Andrius Utka[3] Sigita Rackevičienė[4]

[1]University of Augsburg, Germany [2]University of Cologne, Germany
[3]Vytautas Magnus University, Lithuania [4]Mykolas Romeris University, Lithuania

Languages: Lithuanian - English

Data: cybersecurity domain
1) lexicon (terminology: TBX)
2) corpora (parallel: TMX / annotated: CoNLL)

Challenge: Publish that such that
3) anyone can easily re-use our data
4) we can integrate lexical data, linguistic annotations and parallel corpus
5) we access / query / interlink / process all data with off-the-shelf technology

UniDive: WG2 (mostly)
Status: On-going

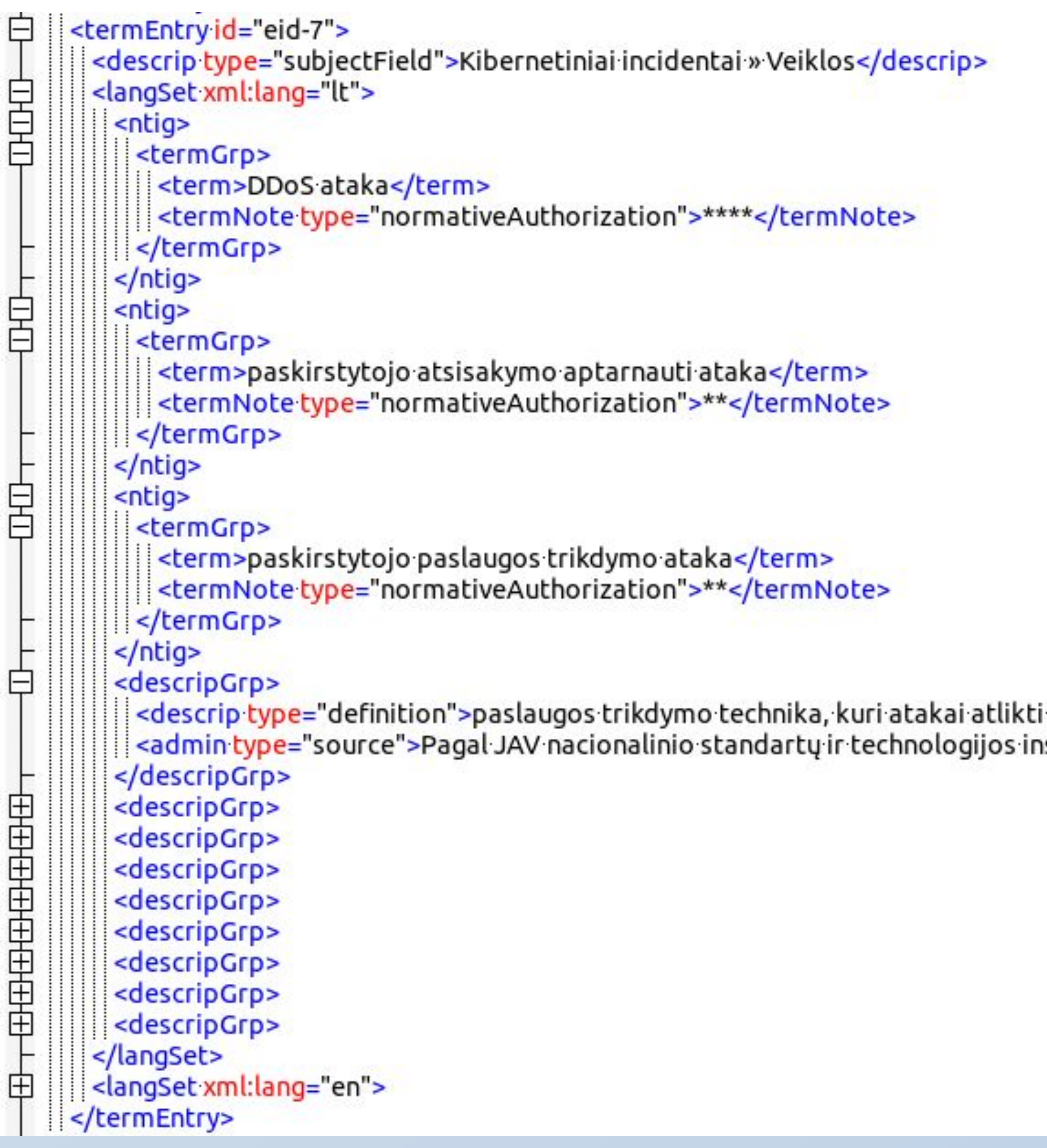
We have a solution that works nicely :)
Yet, aspects of the data modelling are still experimental and need to be supported by additional use cases

TBX: Lithuanian-English Cybersecurity Termbase

- TermBase eXchange Format (TBX v.2)
- 233 cybersecurity concepts with Lithuanian and English designations, definitions and context examples:

  • Concept id
  • Subject field (subdomain)
  • Term group of synonymous terms
  • Description group with definitions
  • Description group with examples

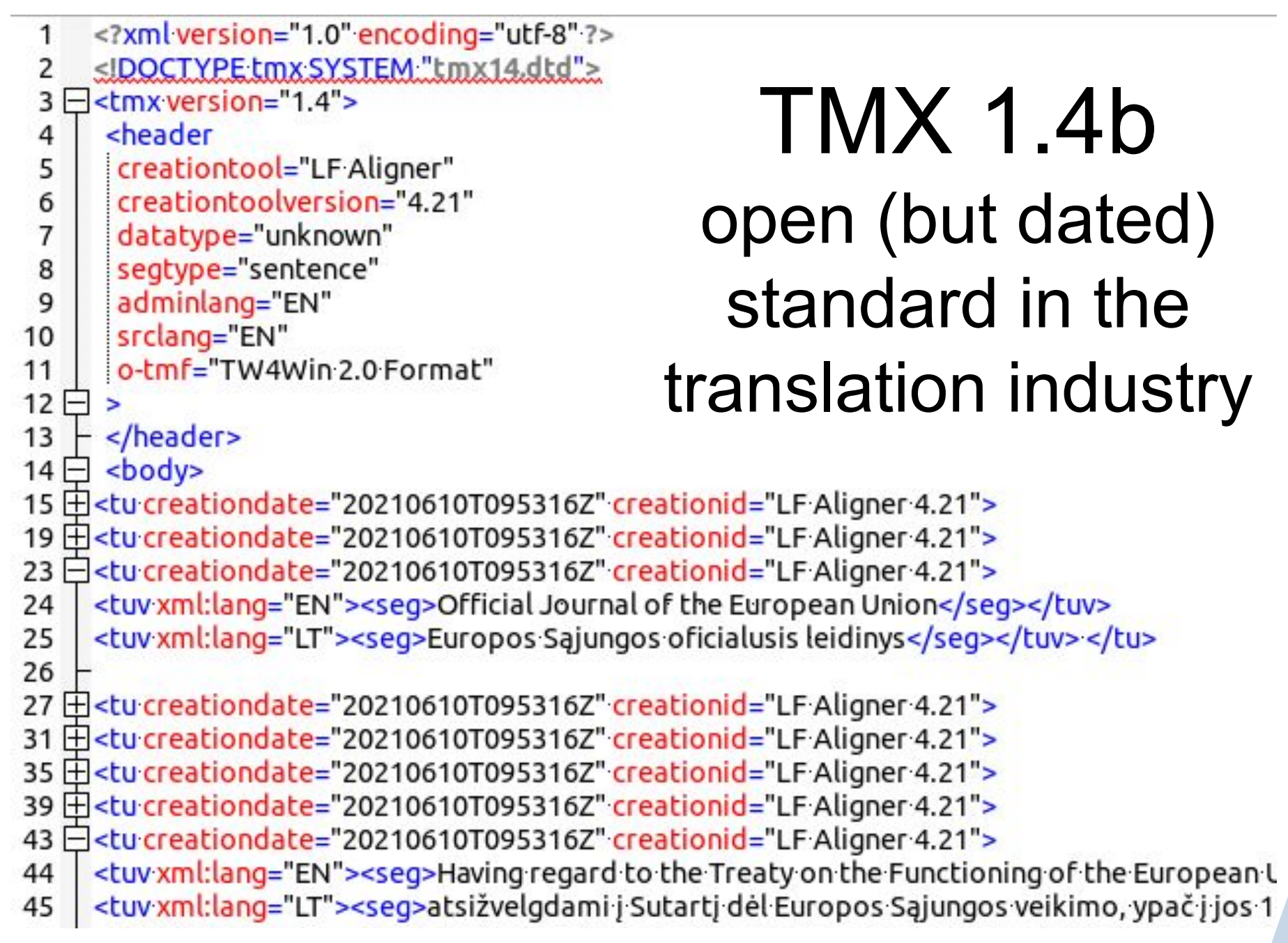
  Data silo, using a domain-specific, proprietary XML standard



```
<termEntry id="eid-7">
  <descrip type="subjectField">Kibernetiniai incidentai » Veiklos</descrip>
  <langSet xml:lang="lt">
    <ntig>
      <termGrp>
        <term>DDoS ataka</term>
        <termNote type="normativeAuthorization">****</termNote>
      </termGrp>
    </ntig>
    <ntig>
      <termGrp>
        <term>paskirstytojo atsisakymo aptarnauti ataka</term>
        <termNote type="normativeAuthorization">**</termNote>
      </termGrp>
    </ntig>
    <ntig>
      <termGrp>
        <term>paskirstytojo paslaugos trikdymo ataka</term>
        <termNote type="normativeAuthorization">**</termNote>
      </termGrp>
    </ntig>
    <descripGrp>
      <descrip type="definition">paslaugos trikdymo technika, kuri atakai atlikti
      <admin type="source">Pagal JAV nacionalinio standartų ir technologijos in
    </descripGrp>
    <descripGrp>
    <descripGrp>
    <descripGrp>
    <descripGrp>
    <descripGrp>
    <descripGrp>
    <descripGrp>
    <descripGrp>
  </langSet>
  <langSet xml:lang="en">
</termEntry>
```

TMX: Parallel Corpus

- Translation Memory eXchange format



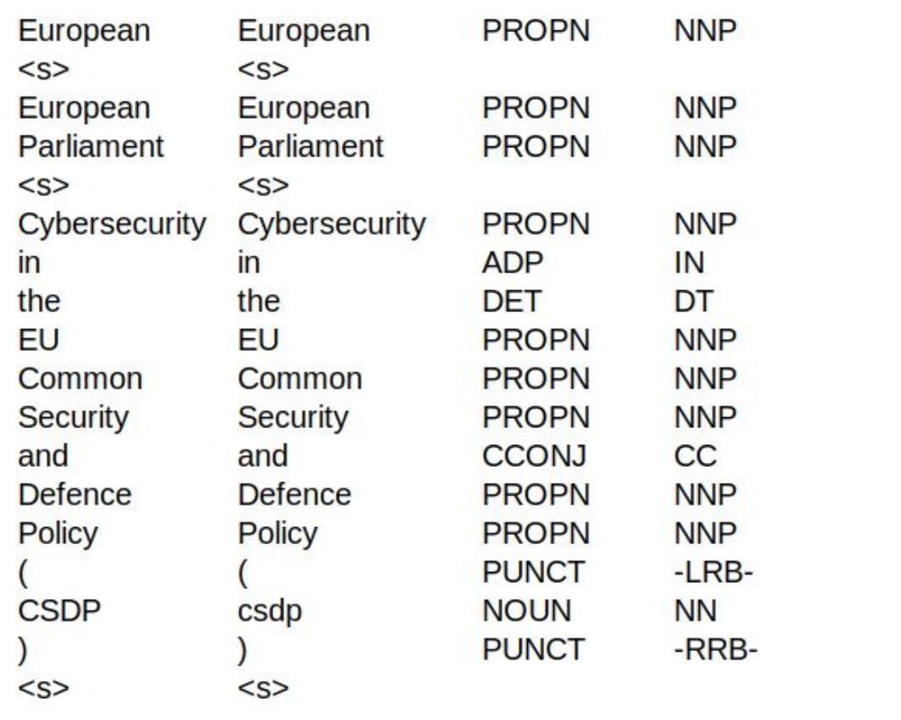
```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
 <header
  creationtool="LF Aligner"
  creationtoolversion="4.21"
  datatype="unknown"
  segtype="sentence"
  adminlang="EN"
  srclang="EN"
  o-tmf="TW4Win 2.0 Format"
 >
 </header>
 <body>
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tuv xml:lang="EN"><seg>Official Journal of the European Union</seg></tuv>
<tuv xml:lang="LT"><seg>Europos Sąjungos oficialusis leidinys</seg></tuv> </tu>

<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tu creationdate="20210610T095316Z" creationid="LF Aligner 4.21">
<tuv xml:lang="EN"><seg>Having regard to the Treaty on the Functioning of the European U
<tuv xml:lang="LT"><seg>atsižvelgdami į Sutartį dėl Europos Sąjungos veikimo, ypač į jos 1
```

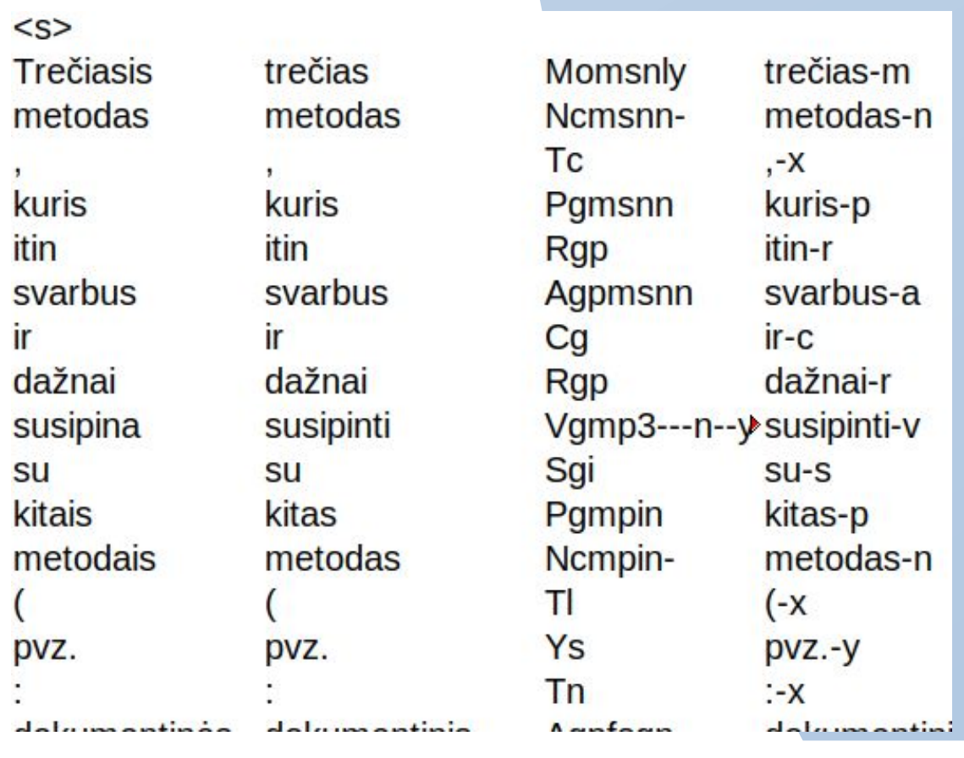
TMX 1.4b
open (but dated)
standard in the
translation industry

Vert: Monolingual Annotations

- CoNLL-style format plus SGML markup
- morphosyntax for both languages



```
European        European        PROPN   NNP
<s>             <s>
European        European        PROPN   NNP
Parliament      Parliament      PROPN   NNP
<s>             <s>
Cybersecurity   Cybersecurity   PROPN   NNP
in              in              ADP     IN
the             the             DET     DT
EU              EU              PROPN   NNP
Common          Common          PROPN   NNP
Security        Security        PROPN   NNP
and             and             CCONJ   CC
Defence         Defence         PROPN   NNP
Policy          Policy          PROPN   NNP
(               (               PUNCT   -LRB-
CSDP            csdp            NOUN    NN
)               )               PUNCT   -RRB-
<s>             <s>
```



```
<s>
Trečiasis       trečias         Momsnly         trečias-m
metodas         metodas         Ncmsnn-         metodas-n
,               ,               Tc              ,-x
kuris           kuris           Pgmsnn          kuris-p
itin            itin            Rgp             itin-r
svarbus         svarbus         Agpmsnn         svarbus-a
ir              ir              Cg              ir-c
dažnai          dažnai          Rgp             dažnai-r
susipina        susipinti       Vgmp3---n--y    susipinti-v
su              su              Sgi             su-s
kitais          kitas           Pgmpin          kitas-p
metodais        metodas         Ncmpin-         metodas-n
(               (               Tl              (-x
pvz.            pvz.            Ys              pvz.-y
:               :               Tn              :-x
```

TMX -> Web Annotation

We are not aware of any precedent for encoding parallel corpora in RDF

Web Annotation is a W3C standard for annotating textual, multimodal and other content on the web with labels or structured information

- We model the translation relation as an annotation that points to two or more “targets”, identified by **XPath Selectors** pointing to the original file
- Additionally, we use the NIF vocabulary to include the **full text** as nif:Context

Vert ->-CoNLL-RDF / NIF

- **CoNLL-RDF** is an established vocabulary and a library for encoding/wrapping TSV data in RDF
- Based on NIF, so, it can be directly **linked to TMX** contexts

TBX -> OntoLex

- **OntoLex**: widely for lexical information in the web of data
  => encode or wrap lexical information in RDF
  => can be flexibly merged with other RDF/RDF-wrapped data, e.g., using the query language SPARQL
- We provide a custom TBX converter converter inspired by Cimiano et al. (2015)

Except for solutions currently pursued in the context of the Linguistic Linked Open Data (LLOD) community, we are not aware of any established technology that allows to seamlessly integrate parallel corpus data, lexical data and morphosyntactic annotations

Conceptually, this requires
- directed graphs => here: **RDF data model**
- community standards for how to encode linguistic information in graphs => here: **LLOD community standards**
- a standard format, designated databases and protocols => here: **RDF/SPARQL** ecosystem

See handout for how to perform cross-lingual queries across corpus and terminology ;)

Universität Augsburg University | UNIVERSITÄT ZU KÖLN | VYTAUTAS MAGNUS UNIVERSITY MCMXXII | MYKOLO ROMERIO UNIVERSITETAS | Nexus Linguarum

christian.chiarcos@uni-a.de
mionov@uni-koeln.de
andrius.utka@vdu.lt
sigita.rackeviciene@mruni.eu