Transforming the Cologne Digital Sanskrit Dictionaries into OntoLex Lemon

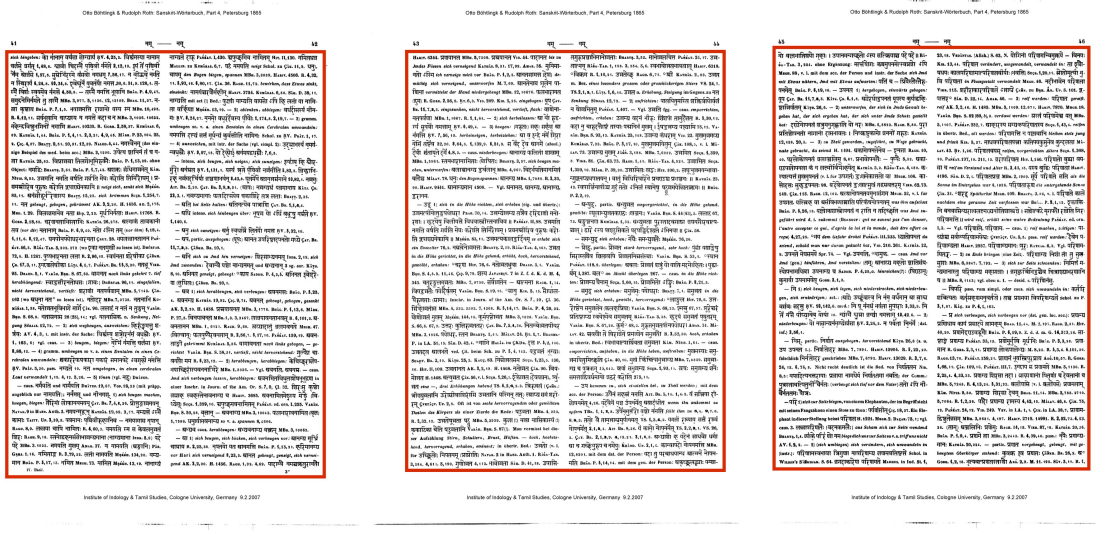
Francisco Mondaca and Felix Rau
CCeH DCH
University of Cologne

The Task: Transforming 36 Dictionary into Ontolex

http://www.sanskrit-lexicon.uni-koeln.de/

36 dictionaries

13	Sanskrit-English dictionaries
3	English-Sanskrit dictionaries
2	Sanskrit-French dictionaries
5	Sanskrit German dictionaries
1	Sanskrit-Latin dictionary
2	Sanskrit-Sanskrit dictionaries
10	specialised (encyclopedic) dictionaries



Sanskrit – German Dictionary
7 volumes
122731 entries

Böhtlingk and Roth Grosses Petersburger Wörterbuch 1855

Sanskrit Lexicography

Complex entries

Typographically dense

अक्षत (3. अ + क्षत) 1) adj. a) *unverletzt* (अहिंसिते) H. an. 3,237. Med. t. 79. दश मासाञ्छयानः कुमारो अधि मातरि । निरैतु जीवो अक्षतो जीवो जीवन्त्या अधि ॥ RV. 5,78,9. अक्षमस्मि सपत्नक्षेन्द्र इवारिष्टो अक्षतः 10,166,2. अजीतो ऽक्षतो अक्षतो ऽध्यस्थां पृथिवीमहम् AV. 12,1,11. दश स्थानानि दण्डस्य मनुः स्वायंभुवो ऽब्रवीत् । त्रिषु वर्णेषु यानि स्युरक्षतो ब्राह्मणो व्रजेत् M. 8,124. राष्ट्रादेनं (ब्राह्मणं) बहिः कुर्यात्समग्रधनमक्षतम् 8,380. अक्षतयोनि 9,176. 10,5. — b) *nicht gemahlen* (अखण्डित) Çabdar. im ÇKDr. अक्षतसक्तूनां नवं कलशं पूरयित्वा Âçv. Gṛhj. 2,1. — 2) n. sg. oder m. pl. *geröstetes Korn* AK. 2,9,47. H. 401. an. 3,238. Med. t. 79. (न द्वयोः) Siddh. K. 249, b, 11. (m. pl.) साक्षतपात्रहस्ता Ragh. 2,21. *Gerste* Med. t. 80 (m. f. n.?). m. = आतपतण्डुल ein Purâṇa im ÇKDr. *Korn* (im Allgemeinen: शस्यमात्रे) Bhânud. zu AK. im ÇKDr. — 3) m. n. *Eunuch* H. an. 3,238. Med. t. 79. — 4) m. Çiva, H. ç. 43. — 5) f. °ता a) *eine unverletzte Jungfrau* eine Smṛti im ÇKDr. — b) Name einer Pflanze = कर्कटशृङ्गी Çabdak. im ÇKDr.

History of the Cologne resources

non-XML markup, ASCII	**1993 – 2001**
XML, Unicode	**2001 – 2012**
XML/TEI (LAZARUS)	**2013 – 2015**
APIs (C-SALT/VedaWeb)	**2017 – 2020**
RDF/Ontolex	**2020 –**

Markup: Typography to semantic

Initial Digitisation: Layout encoded

Entry, headword, language, bold, italics, ...

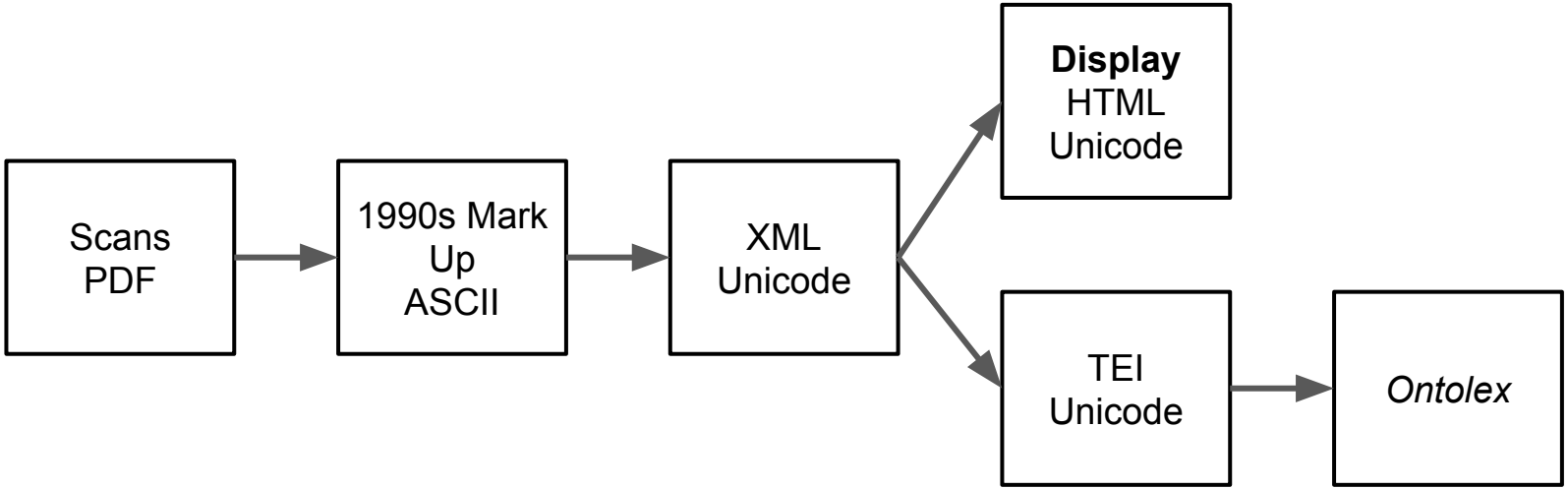
XML: Rough lexicographic structure encoded

added part-of-speech, definition

TEI: Lexicographic microstructures

references, examples, subentry structures, ...

Pipeline


Scans
PDF
1990s Mark
Up
ASCII
XML
Unicode
Display
HTML
Unicode
TEI
Unicode
Ontolex

Sanskrit Lexical Data Accessibility - Status

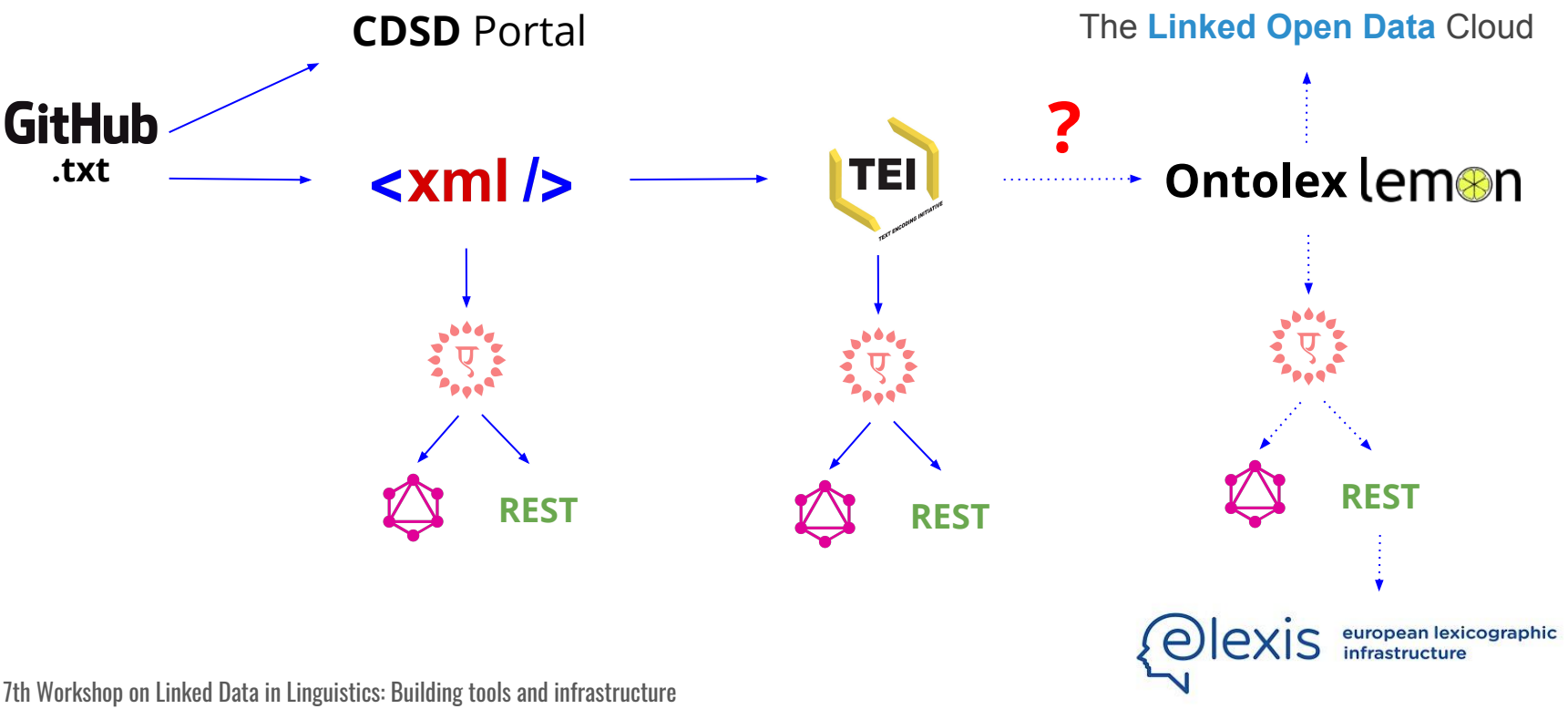
GitHub: .**txt** source (34 dicts).

CDSD Portal: php, JS,SQLite, XML (34 dicts).

Kosh (elasticsearch, REST and GraphQL):

- **XML** → REST and GraphQL APIs (34 dicts) - Sync with CDSD
- **TEI-P5** → REST and GraphQL APIs (7 dicts) - Not sync with CDSD (minimal diff)

Sanskrit Lexical Data Accessibility - Status and Plan



Transforming TEI lexical data into Ontolex-Lemon - I

- 2 encoding methods: RDFa or “pure” RDF.
- Due to the large amount of files to synchronize (36 .txt + 36 XML + 36 TEI), RDFa is the most appealing method.
- Modelling evaluation with the Monier Williams (san-eng) dictionary, one of the most complex of the CDSD collection and the first to be transformed into TEI-P5.

Modelling in Ontolex with RDFa - TEI source

```
<entry ana="H1" xml:id="lemma-aSrAta" xmlns="http://www.tei-c.org/ns/1.0">
    <form>
        <idno ana="hc3">110</idno>
        <orth ana="key1" xml:lang="san-Latn-x-SLP1">aSrAta</orth>
        <idno ana="hc1">1</idno>
        <hyph ana="key2" xml:lang="san-Latn-x-SLP1-headword">a-SrAta</hyph>
    </form>
    <sense>
        <gramGrp>
            <gram ana="lex">mfn.</gram>
        </gramGrp>uncooked
        <cit type="literary_source">
            <bibl xml:lang="san-Latn-x-CSDL">
                <ref target="#auth-RV_">RV.</ref>
                x, 179, 1.</bibl>
        </cit>
        <note>
            <unclear ana="mul"/>
            <idno type="MW">014422</idno>
            <ref target="#page-0114" type="facs">114,2</ref>
            <idno ana="L" xml:id="monier_19802">19802</idno>
        </note>
    </sense>
</entry>
```

Modelling in Ontolex with RDFa - (With errors)

```
<entry typeof="ontolex:LexicalEntry" xml:id="lemma-aSrAta" ana="H1">
    <form property="ontolex:lexicalForm">
        <idno ana="hc3">110</idno>
        <orth property="ontolex:writtenRep" ana="key1" xml:lang="san-Latn-x-SLP1">aSrAta</orth>
        <idno ana="hc1">1</idno
        <hyph property="ontolex:writtenRep" ana="key2" xml:lang="san-Latn-x-SLP1-headword">a-SrAta</hyph>
    </form>
    <sense typeof="ontolex:lexicalSense">
        <gramGrp>
            <gram property="lexinfo:partOfSpeech" ana="lex">mfn.</gram>
        </gramGrp>
        uncooked
        <cit type="literary_source">
            <bibl xml:lang="san-Latn-x-CSDL">
                <ref target="#auth-RV_">RV.</ref>
                x, 179, 1.
            </bibl>
        </cit>
        <note>
            <unclear ana="mul"/>
            <idno type="MW">014422</idno>
            <ref target="#page-0114" type="facs">114,2</ref>
            <idno ana="L" xml:id="monier_19802">19802</idno>
        </note>
    </sense>
</entry>
```

Modelling in Ontolex RDF - I

Encoding complex digitized dictionaries in RDFa requires a deep restructuration of the existing TEI-XML model.

`lexicog` offers data modellers a solution that maps the original structure of a digitized dictionary to Ontolex-Lemon core. This is of great value, specially for dealing with Ontolex-Lemon constrain ‘1 entry = 1 part-of-speech’.

Side effect of this mapping: Increased verbosity and complexity. In some cases: Data redundancy.

Modelling in Ontolex RDF - II

Open questions:

- How to encode references to texts?
- How to encode metadata (original scan, ID in source document)?

धन्यवाद

Slides: https://www.doi.org/10.5281/zenodo.3903138

C-SALT- Cologne South Asian Languages and Texts: c-salt.uni-koeln.de