Universität zu Köln

Introducing C-SALT: Cologne South Asian Languages and Texts

Felix Rau (IfL), Francisco Mondaca (CCeH) | 20.07.18

Overview

- About C-SALT
- Current Projects
  - VedaWeb
  - API for dictionaries

C-SALT

Cologne South Asian Languages and Texts (C-SALT) provides an overview over projects and digital resources related to South Asian languages, texts, and culture at the University of Cologne. C-SALT coordinates the activity of these projects and facilitates sustainable development of the diverse resources.

http://c-salt.uni-koeln.de

- One portal for South Asian resources in Cologne
- Coordination of South Asia related activities
- Shared resources (i.e. expertise, manpower, …)
- Common technological infrastructure

C-SALT

Cologne South Asian Languages and Texts

About C-SALT

Cologne South Asian Languages and Texts (C-SALT) provides an overview over projects and digital resources related to South Asian languages, texts, and culture at the University of Cologne. C-SALT coordinates the activity of these projects and facilitates sustainable development of the diverse resources.

Projects and Resources

Cologne Sanskrit Lexicon

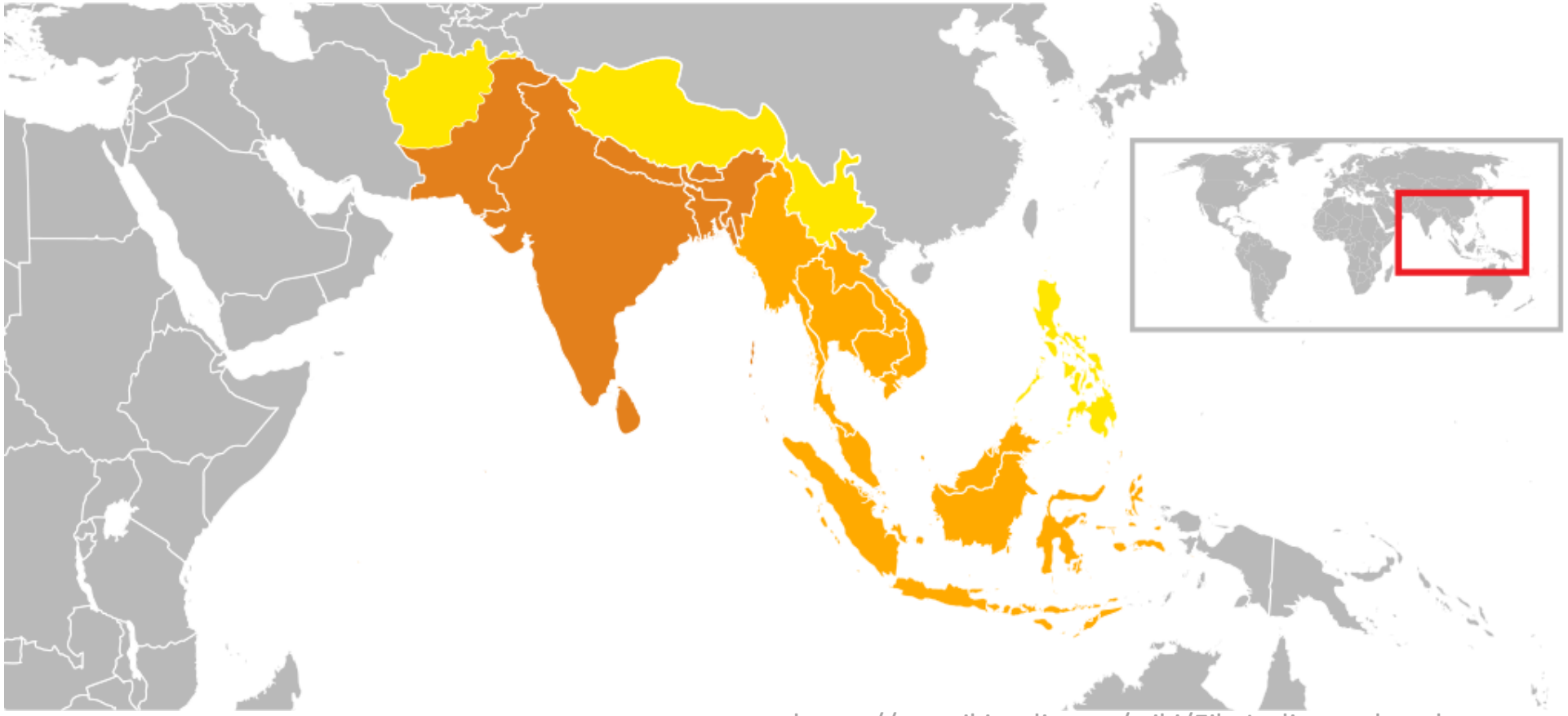
TEI Cologne Sanskrit Lexicon

Critical Pāli Dictionary *Online*

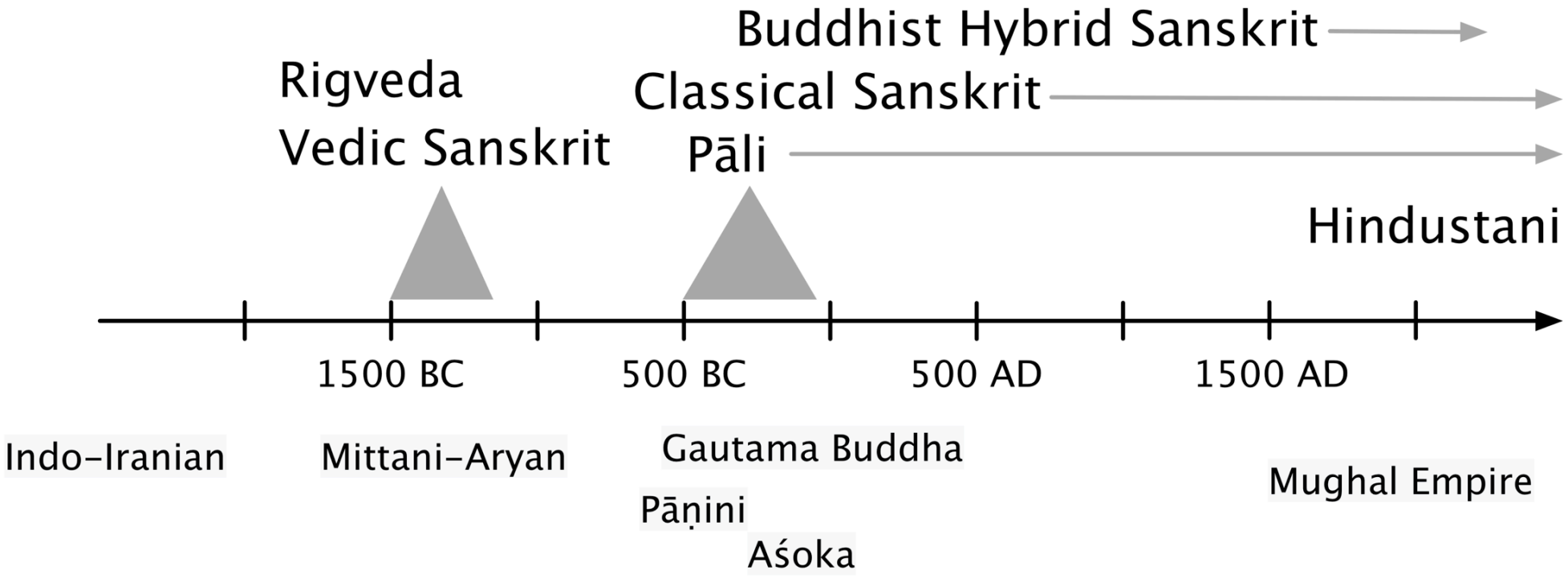
Felix Rau (IfL), Francisco Mondaca (CCeH) | 20.07.18

http://c-salt.uni-koeln.de

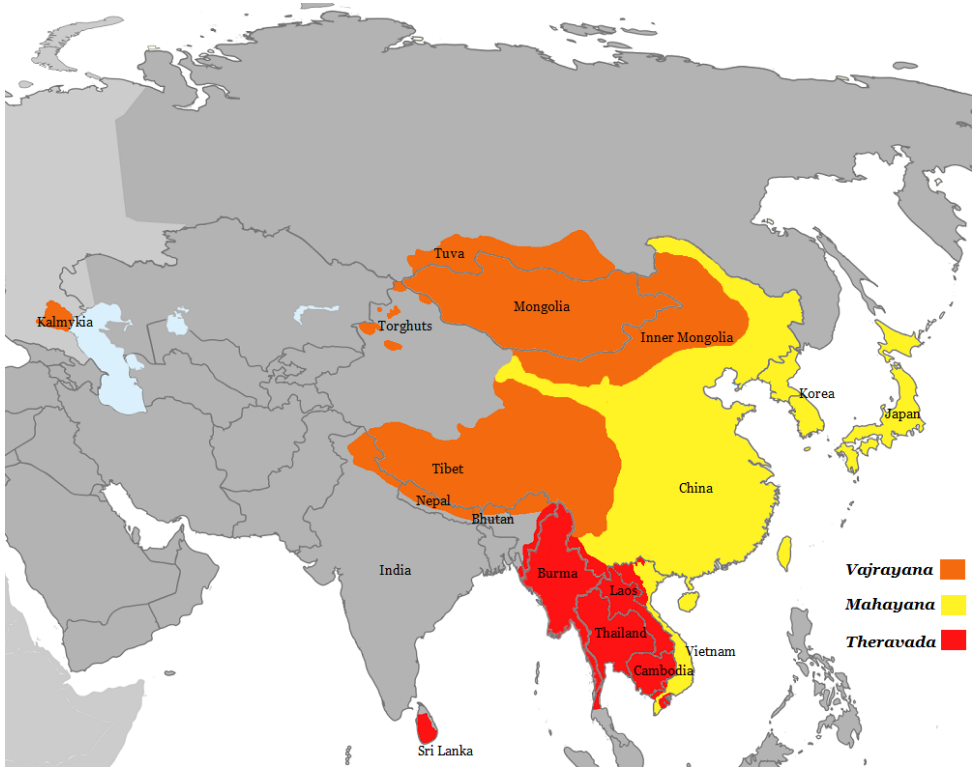
Universität zu Köln



https://en.wikipedia.org/wiki/File:Indian_cultural_zone.svg



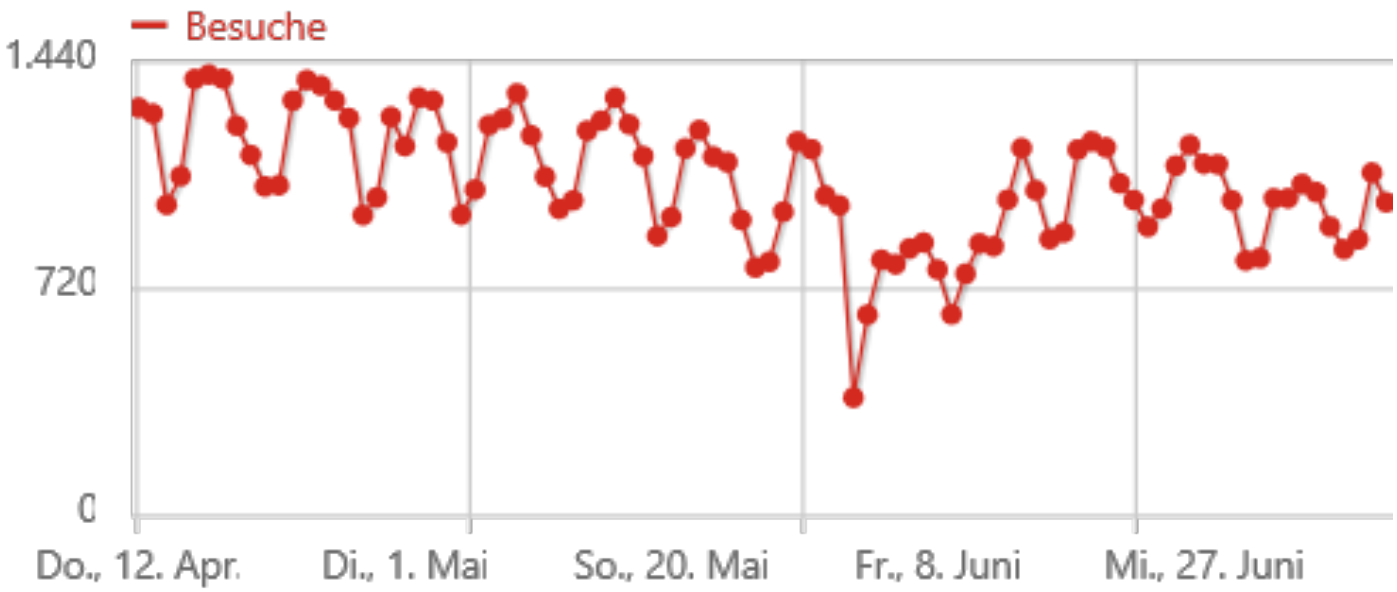
- 448 living languages in India
- 74 living languages in Pakistan
- 121 living languages in Nepal
- 41 living languages in Bangladesh
- …
- Sanskrit liturgical language of Hinduism and Mahāyāna and Vajrayāna Buddhism
  - *India, Sri Lanka, China (esp. Tibet), Mongolia, Japan, Korea, (Indonesia, Vietnam, Russia)*
- Pāli liturgical language of Theravāda Buddhism
  - *Sri Lanka, Burma, Thailand, Cambodia, Laos, (Highland Vietnam)*



C-SALT Resources

- Cologne (Digital) Sanskrit Dictionaries (CSD) (36 dictionaries)
  - 13 Sanskrit-English dictionaries
  - 3 English-Sanskrit dictionaries
  - 2 Sanskrit-French dictionaries
  - 5 Sanskrit-German dictionaries (e.g. PWG 122731 entries)
  - 1 Sanskrit-Latin dictionary
  - 2 Sanskrit-Sanskrit dictionaries
  - 10 specialised (encyclopedic) dictionaries
- TEI Cologne Sanskrit Lexicon
  - 1 Sanskrit-English, 1 Sanskrit-German, 1 English-Sanskrit
- Critical Pāli Dictionary (1924-2010) (Volumes: *A-Ka*, 29726 entries)
- VedaWeb

CSD User Stats 2018



History of the CSD and C-SALT

- 1994 Thomas Malten initiates the project
- 1996 Digitization of the Monier-Williams
- 2003 CSD as a web app
- 2013 Malten retires, the DCH becomes responsible, increasing involvement of the community
- 2013-15 LAZARUS Project (TEI CSL)
- 2016 Critical Pāli Dictionary from Copenhagen
- 2017 VedaWeb
- 2017 C-SALT

C-SALT Development

- Add more Pāli Dictionaries (together with Pāli Text Society)
- Add Iranian dictionaries (and texts)
- Add Nuristani resources
- Add dictionaries from other South Asian languages
- Add texts (e.g. Atharvaveda)
- Cross dictionary search for CSL and Pāli Dictionaries
- APIs to provide dictionaries for other projects and services

About VedaWeb

- Web platform for Vedic Sanskrit Texts
- DFG-funded project (2017-2020)
- Vedic Sanskrit:
  - Ancient Indo-European language from the northern Indian subcontinent
  - Language of the Vedas (Rigveda, Yajurveda, Samaveda, Atharvaveda), religious texts (veda: knowledge) that comprise the oldest layer of Sanskrit literature
- Cornerstone of the project: Rigveda

About the Rigveda

- One of the oldest and most of important texts of the Indo-European language family (~ 1500-1000 B.C.).
- Oldest of the main texts of hinduism.
- Larger than the Iliad and Odyssey combined:
  - 10 books (maṇḍalas)
  - 1.028 hymns (sūktas)
  - 10.552 verses
  - 164.766 tokens
  - 32.131 types
- Due to its origin and structure is a corpus of great research interest.

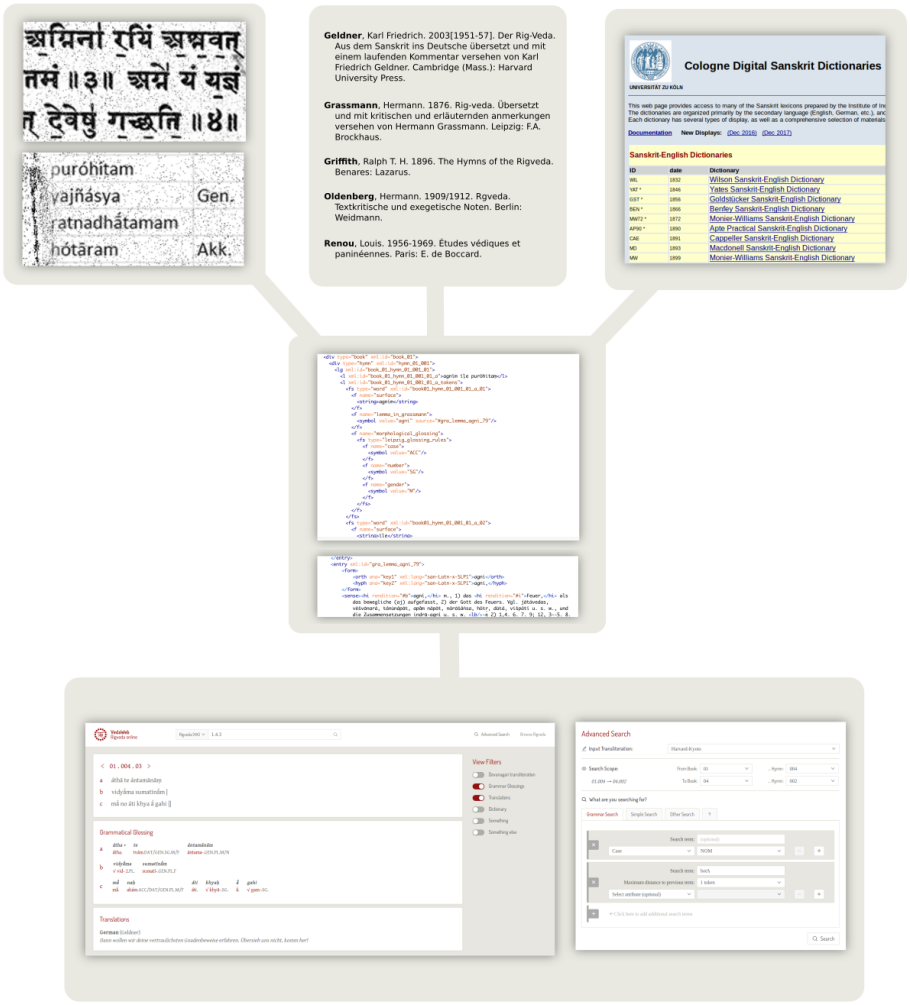
Rigveda in VedaWeb: Sources I

- Basis for VedaWeb is a version of the Rigveda annotated with morpho-syntactic information by Paul Widmer & Salvatore Scarlata, University of Zurich.
- Metrical information by Dieter Gunkel (University of Richmond) & Kevin Ryan (University of Harvard).
- Morpho-lexical, Verb-Argument annotations by Oliver Hellwig (University of Düsseldorf), Heinrich Hettrich (University of Würzburg) et al.

Rigveda in VedaWeb: Sources II

- Hermann Grassmann (1809-1877) compiled a dictionary specially for the Rigveda. Each token in the corpus (Zurich) is linked to an entry in Grassmann's dictionary*. This an other dictionaries are to be linked with the text.
- There are different translations (EN, DE, FR) of the Rigveda. Some of them will be included in the project.

Rigveda in VedaWeb: Workflow



Felix Rau (IfL), Francisco Mondaca (CCeH) | 20.07.18

Universität zu Köln

Rigveda in VedaWeb: TEI (Text Encoding Initiative) Modelling

- Why TEI?
  - TEI is a collectively developed standard used for representations of texts in digital form.
  - De facto standard in Digital Humanities projects dealing with texts.
  - Different Special Interest Groups (SIGs) covering multiple research areas.
  - Relevance for our project: provides a consistent data model and ensures **data persistence**.

Rigveda in VedaWeb: TEI Model, Web app

- Current status

Rigveda in VedaWeb: Problems

- Each token in the Rigveda (University of Zurich) is linked to an entry in Grassmann's dictionary. Referenced lemmas in Zurich's file sometimes does not relate 1:1 to the lemmas in Grassmann's dictionary -> Disambiguation is required.

Cologne Sanskrit Dictionaries (CSD) – in TEI

- Lazarus Project (2013-2015):
  - Monier-Williams (1899) [SA-EN]
  - Böthlingk-Roth (1855-1875) [SA-DE]
  - Apte‘s Student (1920[3)] [SA-EN]
- Soon to be published in TEI format:
  - Grassmann (1873) [SA-DE]
  - The Vedic Index of Names and Subjects (1912) [SA-EN]
  - Apte Practical Sanskrit-English Dictionary (1890) [SA-EN]
  - Böhtlingk Sanskrit-Wörterbuch in kürzerer Fassung (1879) [SA-DE]
  - Stchoupak Dictionnaire Sanscrit-Français (1932) [SA-FR]

Sanskrit transliterations - I

- Most of the texts part of the Cologne Sanskrit Dictionaries are printed in Latin script. But diacritics used in these texts:
  - Were impossible to encode with ASCII (American Standard Code for Information Exchange)
  - Are hard to type with a Western keyboard
- For these reasons scholars developed their own ASCII-based transliterations for Indic script (Devanagari), e.g.:
  - Harvard-Kyoto
  - SLP1 (Sanskrit Library Phonetic Basic encoding scheme)
- Since 2001: ISO 15919 - "Transliteration of Devanagari and related Indic scripts into Latin characters". Codifies stablished practices among Indologists.

Sanskrit transliterations – II

Devanagari	हिरण्यवी
SLP1	hiraRyavI
Harvard-Kyoto	hiraNyavI
ISO 15919	hiraṇyavī
Grassmann	hiraṇya-vî

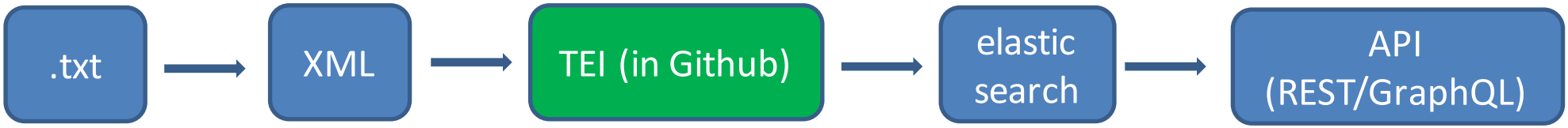
Why TEI for Cologne Sanskrit Dictionaries (CSD)?

- Digitization/transcription of the CSD was a great achievement when done years ago (started 1996) – XML and Unicode were not available! But this process is deprecated regarding the special encoding and markup employed.
- First step of this project is to adapt all dictionaries to a TEI schema already developed during the Lazarus project and encode the texts with ISO 15919.
- TEI and ISO 15919 as standards secure resource sharing and their persistence.

An API (Application Programming Interface), why?

- Most of the dictionaries available online are monolithic applications (frontend + backend all-in-one)
- For a project such as VedaWeb, where lemmas are to be referenced programmatically, an API is necessary, specially if multiple dictionaries have to be linked.
- Current possibilities: make the dictionaries available either as a REST API or a GraphQL API

API - Pipeline



- All TEI dictionaries will be hosted in Github. If the content of the dictionaries is modified (typos, etc.), these changes can be easily and openly tracked.
- Every commit affecting the content of the dictionaries will trigger an automatic update of the index (in elastic search).

धन्यवाद