

The Kosh Suite: A Framework for Searching and Retrieving Lexical Data Using APIs

Francisco Mondaca, Philip Schildkamp, Felix Rau,
Luke Günther
University of Cologne



Overview

- What is Kosh?
- The rationale behind Kosh
- How Kosh works
 - Preparing your data
 - Deployment options
 - Building search queries
- Complementary tools
- Use cases

Kosh

... is a framework to serve diverse lexicographic data

- flexibly
- sustainably
- with minimal configuration

Kosh provides APIs for XML-encoded dictionaries

- format agnostic
- serves REST and GraphQL APIs

... but also comes with a default user interface

```

kosh.uni-koeln.de/ap/ducange/restful/entries?field=
https://kosh.uni-koeln.de/ap/ducange/restful/entries?field=
JSON Raw Data Headers
Save Copy Collapse All Expand All Filter JSON
▼ data:
▼ entries:
▼ 0:
  created: "2023-06-01T13:44:01.997850"
  xml:
    <entry xmlns="http://www.telc-c.org/ns/1.0" xml:id="SECHHARIA" rend="bene
    rend="b">SECHHARIA</form>, i. au Italico <hi rend="i">Seccia</hi>, Rejcti
    proficentis immunditas in viis. Iden intelligitur devn
    lemma: "SECHHARIA"
    quote_lat:
      0: "De pona proficentis immunditas in viis. Iden intelligitur devn
      id: "SECHHARIA"
  1:
    created: "2023-06-01T13:42:43.254111"
    xml:
      <entry xmlns="http://www.telc-c.org/ns/1.0" rend="carpentier" xml:id="AFFA
      xml:id="AFFAITEMTUQ2-1"><form rend="b">num=2</num>, <num> AFFAITEMTUQ2/fo
      Stat. Turin, ann. 1949, cap. 94, ex Cod. reg. 4822, A. : <quote xml:lang="
      Affaitmentore, vel incturas in viis publicis. Affaitmentore, <quote-
      rend="sc">Affaitum</form>, Eadem notione, Stat. civit. Saluciar. collat. 3
      in viis publicis dicta civitatis Saluciarum.</quote> Stat. Av
      ponere vel poni facere in viis vel plateis publicis infra  — hu
      Vide mox <ref chRef="AFFAITARE">Affaitare</ref>, et <ref target="AFFEITARE"
    lemma: "AFFAITEMTUQ2"
    rend_sci:
      0: "Affaitum"
    ref_target:
      0: "AFFEITARE"
    ref_chRef:
      0: "AFFAITARE"
    quote_lat:
      0: "Eadem\N ponere sustinet quilibet affaitator, inctur, vel pel
      1: "Idem bannum\N solvant calligarii; seu alii poneses ruscatum
      2: "Nullum persona possit, vel debeat ponere vel poni facere in viis vel plateis publicis infra\N burgo aliquam ruscam Affaiti, vel aliquas pellaturas coriorum
      id: "AFFAITEMTUQ2"
  2:
    created: "2023-06-01T13:42:57.065379"
    xml:
      <entry xmlns="http://www.telc-c.org/ns/1.0" xml:id="CALLIGARIUS" rend="carpentier"><form=orth type="lemma">CALLIGARIUS</orth>=</form><dictScrap xml:id="CALLIGARIUS-1"><form
      rend="b">CALLIGARIUS</form>, <hi rend="i">Calligarus</hi> confector, Stat. Saluc. collat. 3, cap. 92 : <quote xml:lang="lat">Idem bannum\N solvant Calligarii seu alii poneses
      ruscatum affaiti in viis publicis.</quote> Vide\N <ref chRef="CALLIGARIUS">Calligarus</ref> in <ref target="CALLIGA">Calliga</ref>.</dictScrap></entry>
    lemma: "CALLIGARIUS"
    ref_target:
      0:

```

Kosh - APIs for Lexical Data

Collection: C-SALT Sanskrit Dictionaries: ae ap90 bhs gra mw

Field: headword_slp1 Query Type: fuzzy Query Size: 20

agnī Search

Search Results

Display fields: created headword_deva headword_iso headword_slp1 id

headword_deva	headword_iso
अग्नि	agnī
अग्निम्	agnidh

VedaWeb Rigveda online

Quick search (HK) in "All Text Versions"

Search in: All Text Versions

RegEx:

Input method: HK (Harvard-Kyoto)

Accent-sensitive:

01 001 01

Toggle content Full size view Export

Text Versions

agnīm iḷe purōhītam yajñāsya devām ṛtvijam hōtārah ratnadhātamam
अग्निमीळे पुरोहितं यज्ञस्य देवमुक्त्विभम् होतारं रत्नधातमम्

Dictionaries

Lemma	Full Forms	Preview (Graßmann)	Entries (Graßmann)	Others
agnī-	(agnīm)	agnī,m., (1) das Feuer, als das bewegliche (aj) aufgefasst, (2) der Gott des Feuers. V...	#1	n/a
viḷ- ~ viḷ-	(iḷe)	īd(neutr. des Deutestammes i) hebt den durch das vorhergehende (betonte) Wort ...	#1 #2 #3	n/a
devā-	(devām)	devā,a., m. (das f. devī siehe für sich), (1) a., himmlisch [von div]; insbesondere wi...	#1	n/a
purōhīta-	(purōhitam)	purōhīta,a., m. [ursprünglich Part. von dhā mit purás],1) a., einem Werke [D. L.] v...	#1	n/a
yajñā-	(yajñāsya)	yajñā,m., bisweilen yajanā zu lesen, Götterverehrung, die Reihe der Handlungen, ...	#1	n/a

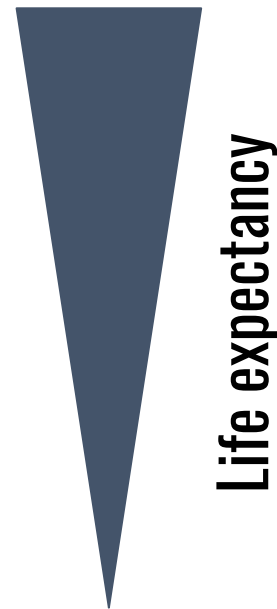
Motivation

- We are supporting a variety of projects at the University of Cologne and beyond
- We might have limited influence on data models and serialization
- We are responsible for sustainably providing access to lexicographic data, even after the project has ended
- Project require different types of data integration and data presentation

→ Focus on generic and long lasting parts of the infrastructure

Infrastructure Life Expectancy

- Data models
- Data serialization
- API definitions
- Backend software
- Frontend software



Focus: Providing APIs

- Actively serve data in an easily usable way (not just preserve it in a repository)
- Accommodate different data models and data serializations
- The backend software can be refactored or rewritten
- Data can be integrated in project specific presentations and applications
- Data can be used and viewed even if project specific solutions can no longer be maintained

Most user-friendly, sustainable investment of work for our situation

VedaWeb | Stanza 1.1.1 | Hymns x +

https://vedaweb.uni-koeln.de/rigveda/view/index/0

VedaWeb
Rigveda online

Quick search (HK) in "All Text Versions"

Search in: All Text Versions

Regex:

Input method: HK (Harvard-Kyoto)

Accent-sensitive:

Browse Rigveda | Advanced Search | About VedaWeb | Guided Tour | Help & Instructions

< 01 . 001 . 01 >

Toggle content | Full size view | Export

Text Versions

agním iḷe puróhitaṃ yajñásya devám ṛtvjám hótāraṃ ratnadhātamaṃ
अग्निमीळे पुरोहितं यज्ञस्य देवमृत्विजम् होतारं रत्नधातमम्

Dictionaries

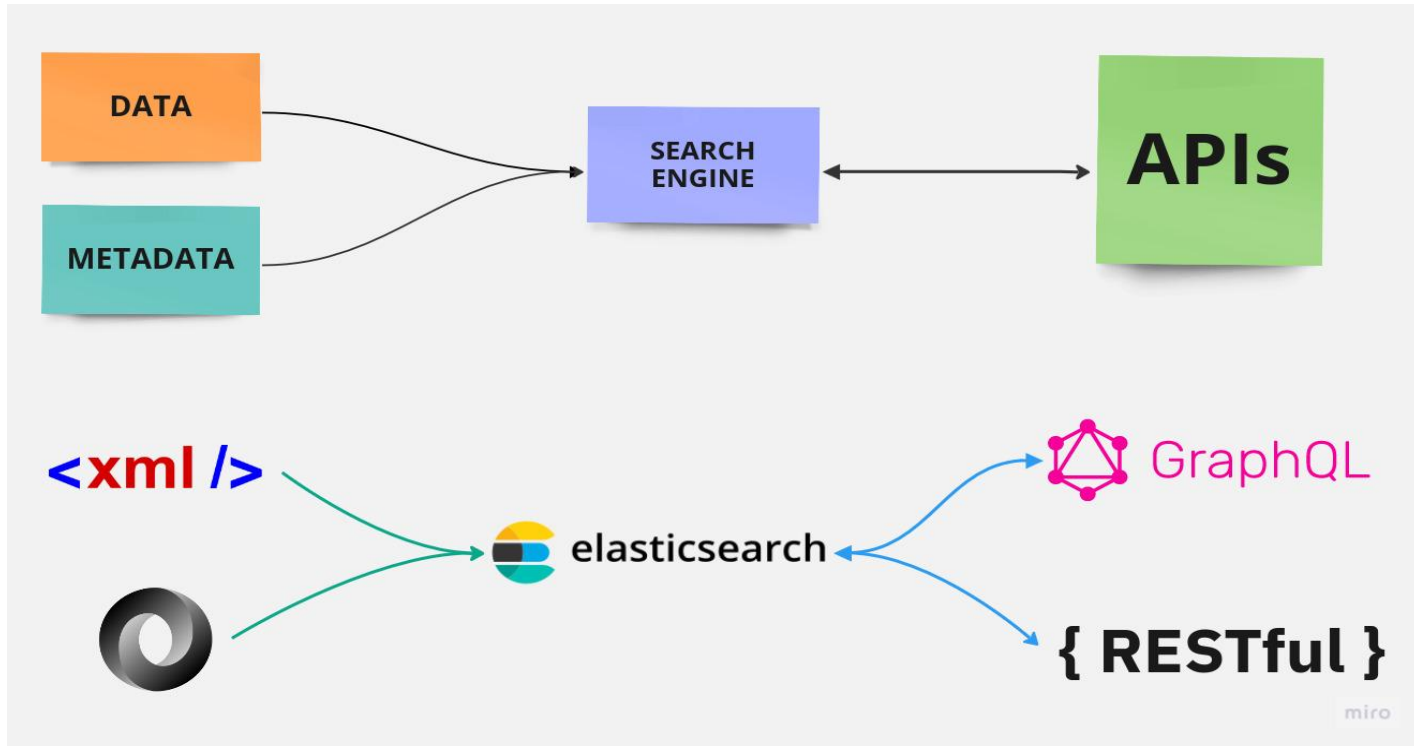
Lemma	Full Forms	Preview (Graßmann)	Entries (Graßmann)	Others
agní-	(agním)	agní,m., (1) das Feuer, als das bewegliche (aj) aufgefasst, (2) der Gott des Feuers. V...	#1	n/a
vīḍ- ~ vīḷ-	(īḷe)	íd[neutr. des Deutestammes i] hebt den durch das vorhergehende (betonte) Wort ...	#1 #2 #3	n/a
devá-	(devám)	devá,a., m. (das f. devī siehe für sich), (1) a., himmlisch [von dív]; insbesondere wi...	#1	n/a
puróhita-	(puróhitam)	puróhita,a., m. [ursprünglich Part. von dhā mit purás],1) a., einem Werke [D. L.] v...	#1	n/a
yajñá-	(yajñásya)	yajñá,m., bisweilen yajaná zu lesen, Götterverehrung, die Reihe der Handlungen, ...	#1	n/a
ratnadhātama-	(ratnadhātamaṃ)	rátna n. (m. 460.10. Gabe, Schatz, Reichtum, Gut als geschenktes [von rā], vel. m...	#1	n/a

<https://vedaweb.uni-koeln.de/>

Kosh Suite: How it works



Kosh Suite: How it works



Kosh Backend: Building blocks

- Kosh can be used as is with any data already existing in XML format
- All you need are two simple configuration files

First, you specify how your data should be **indexed**.

→ Mapping XML nodes to search fields

Second, you add any **metadata** fields you want to be available.

→ Represented languages, authors/contributors, timeframe, licenses

Mapping XML Data

```
<entry id="13">
  <form>
    <orth>abadetasun</orth>
  </form>
  <sense n="1">
    <gramGrp>
      <pos>
        <q>iz.</q>
      </pos>
    </gramGrp>
    <def>monasterioko buruaren kargua eta egitekoa</def>
  </sense>
  <sense n="2">
    <gramGrp>
      <pos>
        <q>iz.</q>
      </pos>
    </gramGrp>
    <def>apaizgoa</def>
    <usg type="geo">
      <q>Bizk.</q>
    </usg>
  </sense>
</entry>
```

```
{
  "mappings": {
    "meta": {
      "_xpath": {
        "id": "./@id",
        "root": "//entry",
        "fields": {
          "lemma": "./form/orth",
          "[sense_def]": "./sense/def",
          "[sense_pos]": "./sense/gramGrp/pos/q",
          "[dicteg]": "./sense/dicteg/q"
        }
      }
    },
    "properties": {
      "lemma": {
        "type": "keyword"
      },
      "sense_def": {
        "type": "text"
      },
      "sense_pos": {
        "type": "text"
      },
      "dicteg": {
        "type": "text"
      }
    }
  }
}
```

How to serve data with Kosh

Currently, there are three ways you can use Kosh:

- Our demo instances
- Docker container
- Native support on Unix-like systems → Just grab the code from GitHub!

Real-time updates: Once Kosh has been started, it will watch your data directory for changes and will continuously index new items.

Kosh sample data: Explore Kosh by looking at both external historical but also current project data from our institutes in Cologne.

Querying Data

- Search strings are matched against the content of a specific property for each lexical item in your database
- There are various search strategies, e.g. fuzzy matching or wildcard search

You can inspect and test the two API endpoints through graphical interfaces!

A typical query for the **REST API** consists of the **query string**, the **field to be queried** and the **search strategy** and, optionally, the **query size**:

```
https://<endpoint>/<dictionary_id>/restful/entries?query=search%20query&field=xml&query_type=match&size=30
```

Complementary tools: Client and Sync

The **Kosh Client** enables you to:

- Compare dictionaries with each other in a familiar table layout
- Search multiple different collections of dictionaries
- Create granular queries and customize the way results are displayed

→ Can be used locally but the Kosh backend has to be reachable first.

The **Kosh Sync** service can be run additionally if you want continuous integration and deployment of your data from a Git repository.

Summary: Use cases for the Kosh Suite

Kosh can help you:

- Get an overview of your current work
- Gauge how far you have come in your data selection and curation process
- Check for errors in your data
- Provide easy access for fellow researchers
- Present your research data to the public, e.g. on a project website

Getting help

For further information, please consult the documentation at:

<https://kosh.uni-koeln.de>

Or send an email to:

info-kosh@uni-koeln.de

We're happy to help!

Thank you for your attention!

