Software Architecture for the Automatization of Subject Indexing

Christopher Bartz,
ZBW – Leibniz Information Centre for Economics
Riga, Latvia 08.06.2022

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

The ZBW is a member of the Leibniz Association.

Overview

- Motivation
- Architecture
- Technologies
- Infrastructure

Motivation

Subject Indexing @ ZBW

Document

World / Domain Knowledge

Meta data

ID1, ID3, ID4

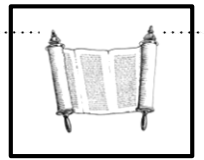
STW
ID1 …
ID2 …
ID3 …
…

Controlled vocabulary

Discovery System

ECONBIZ
Find Economic Literature.

Search:



Title: …
Author: …
Subjects: ID1, ID3, ID4

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

https://www.econbiz.de/

Need for Automation

- > 100 000 new publications each year
- Human Subject Indexing at ZBW can only index ~ 35000 a year
- Few opportunities to reuse third-party data
- Automation can assist subject indexers with suggestions

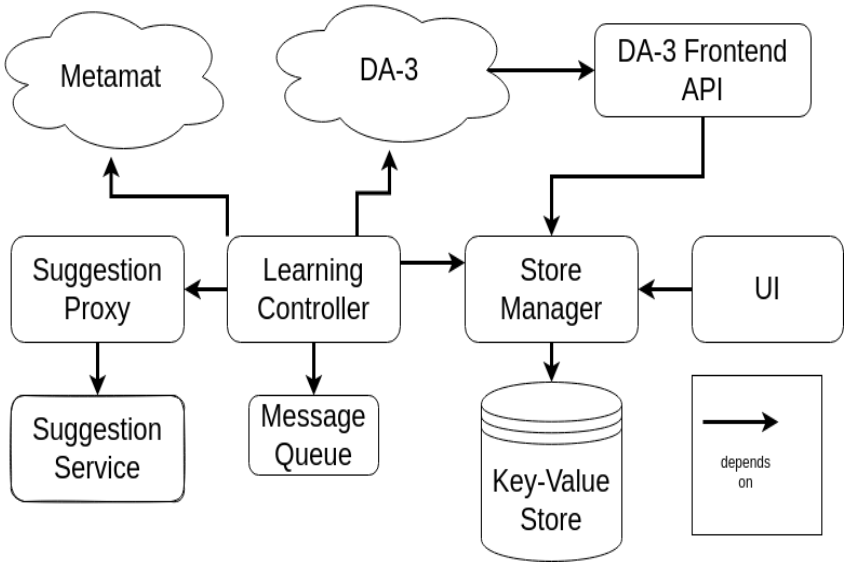
Road to Now

- Started automated indexing projects in 2002
- Since 2014 with inhouse PhD / research engineer
- Since 2020 with the goal to transfer results into a productive service
- Team
    - Dr. Anna Kasprzik – Team Lead
    - Moritz Fuerneisen – Research Engineer
    - Christopher Bartz – Software Developer

Software Architecture

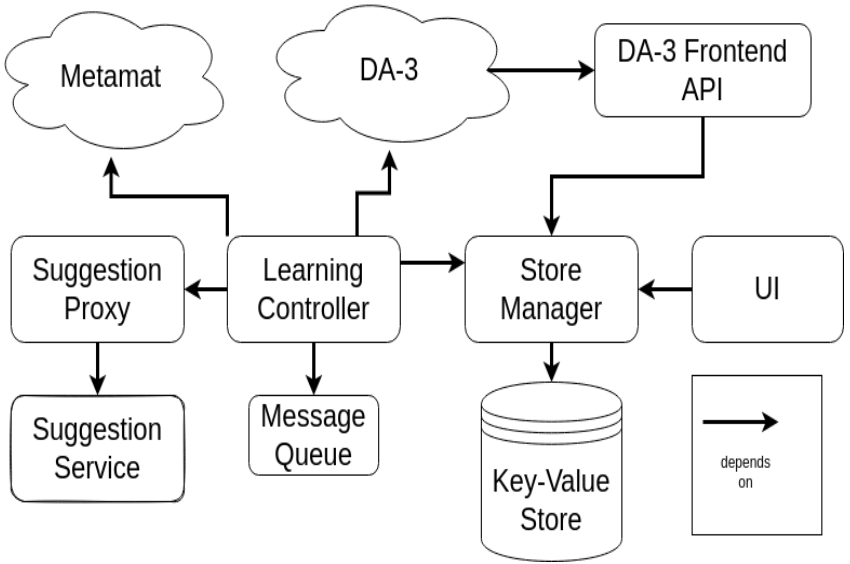
Software Architecture – Overview

- Microservices
- Loose coupling
- REST APIs
- HTTP



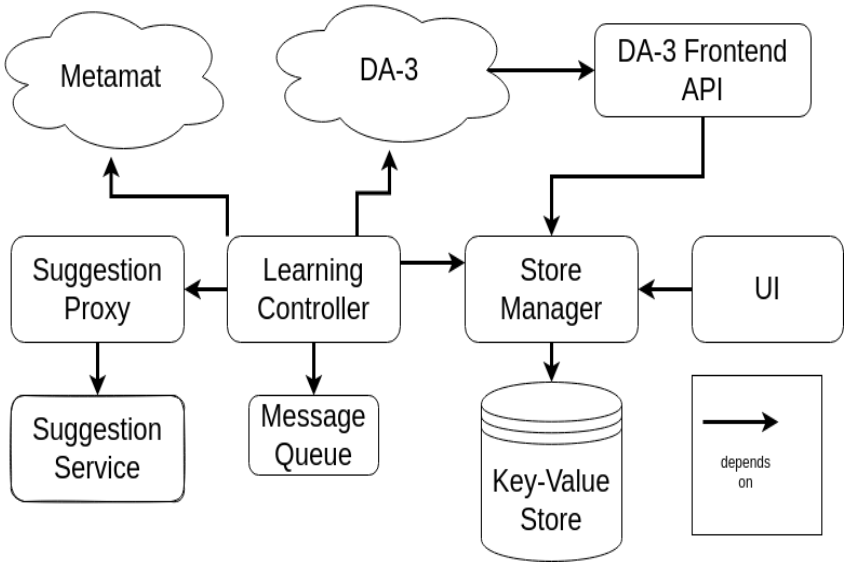
Software Architecture – Components I

- Store Manager: Encapsulates access to K/V Store
- DA3-Frontend: Fetches suggestions from store by request of DA-3
- Learning Controller: Creates objects from external sources and stores them in Store

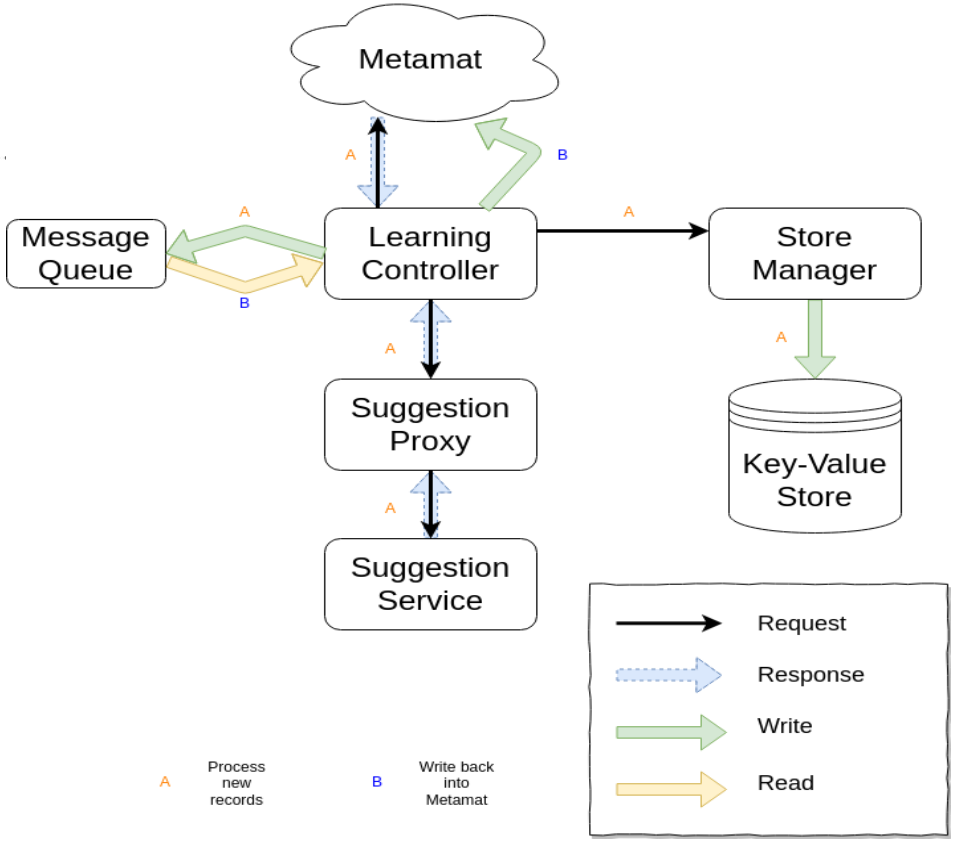


Software Architecture – Components II

- Suggestion Service: Generates suggestions for publications
- Suggestion Proxy: Manages access to Suggestion Service and applies filter & quality rules
- UI: Displays statistics



Interaction with Metamat



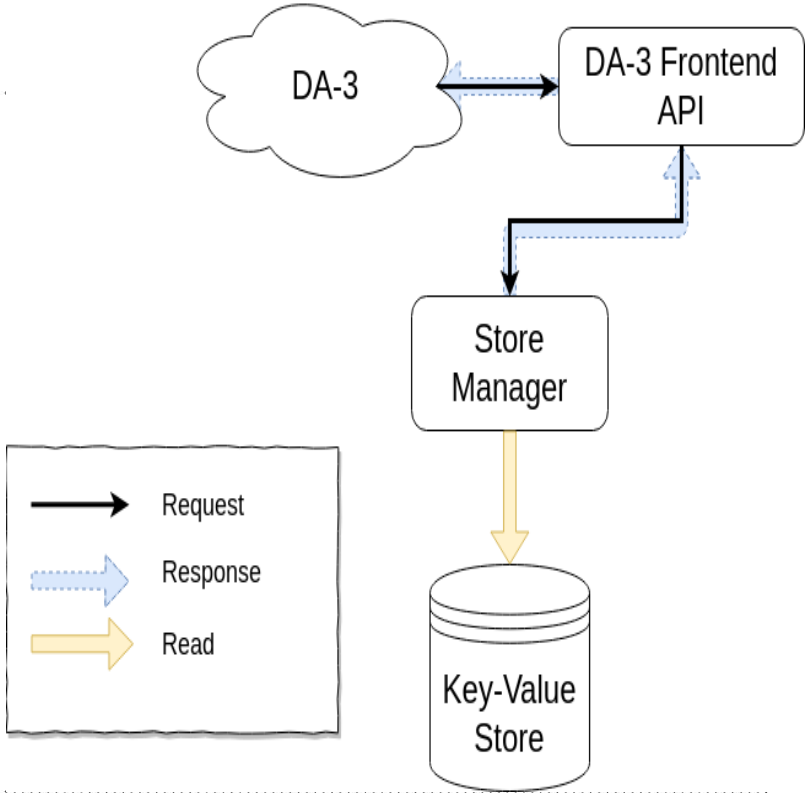
Econbiz

Signature experience : art and science of customer engagement for fashion and luxury companies

edited by Stefania Saviolo

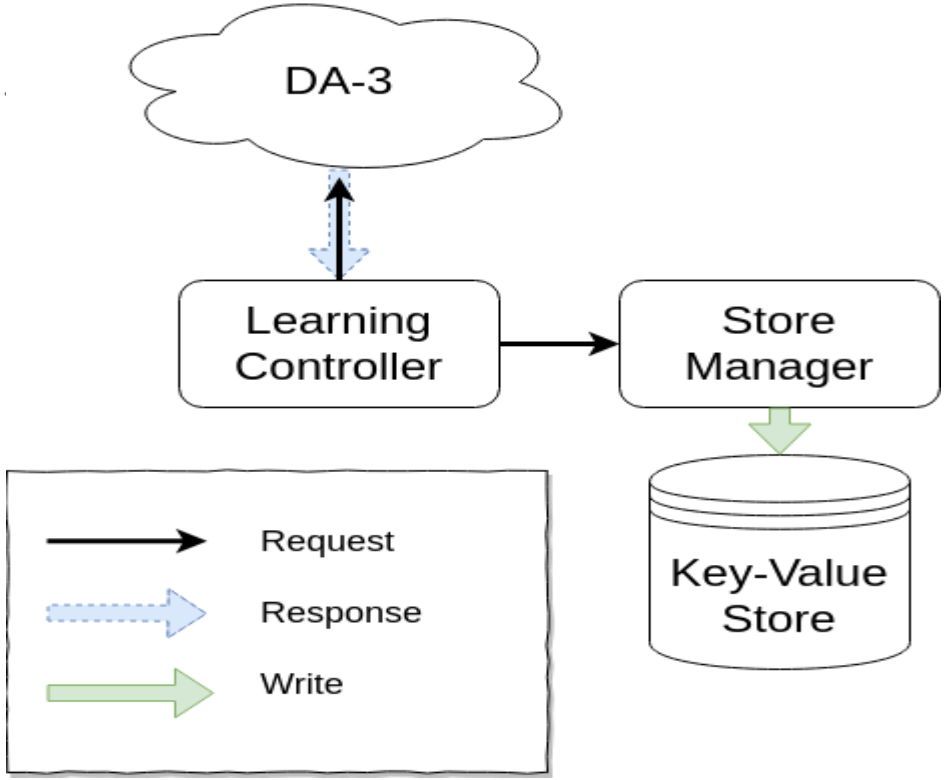
Year of publication:	August 2018 ; First edition
Other Persons:	Saviolo, Stefania (ed.)
Publisher:	Milano : BUP
Subject:	Luxusgüter | Luxury goods | Mode | Fashion | Markenführung | Brand management | Beziehungsmarketing | Relationship marketing | Konsumentenverhalten | Consumer behaviour
Description of contents:	Table of Contents [gbv.de]

Suggestions in DA-3



Retrieval of Assessments

Tools 〉 **Bewertung**		Einstellungen #
Bewertung abschicken		6/6
Gesamtbewertung		
Quelle zbwase		++ + o - | ×
STW		
Beziehungsmarketing	zbwase	++ + o - | ×
Konsumentenverhalten	zbwase	++ + o - | ×
Luxusgüter	zbwase	++ + o - | ×
Markenführung	zbwase	++ + o - | ×
Mode	zbwase	++ + o - | ×



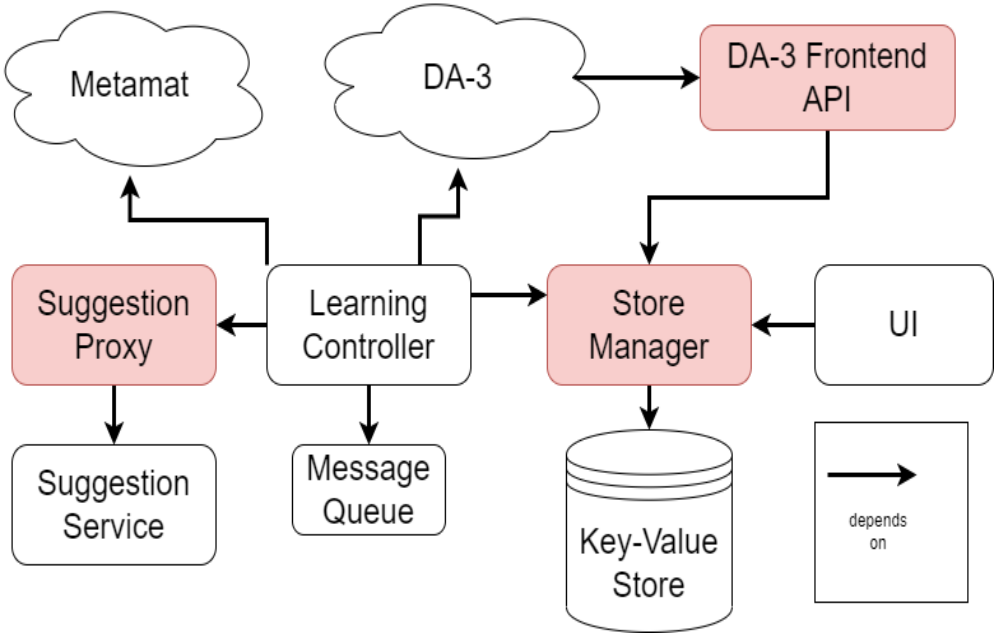
Technologies

Requirements

- Python compatible
- GPL compatible license
- Lightweight
- Continuously maintained
- Acceptable community size

REST APIs

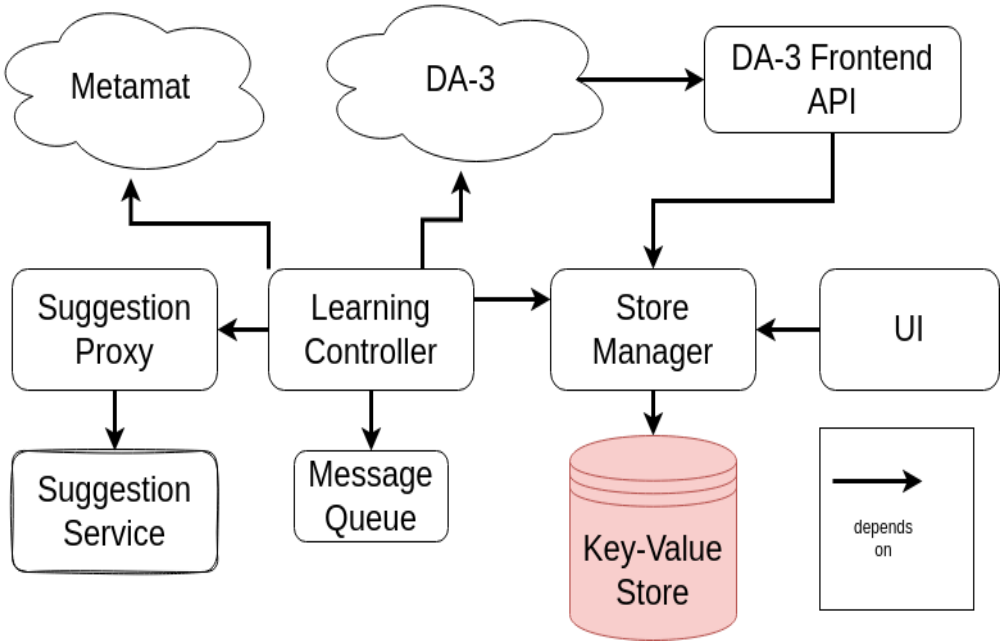
- Web Framework FastAPI
  - automated validation, serialization, documentation, OpenAPI spec generation
  - Swagger UI
- OpenAPI Client Generator
- JSON Format



https://fastapi.tiangolo.com
https://openapi-generator.tech

CouchDB

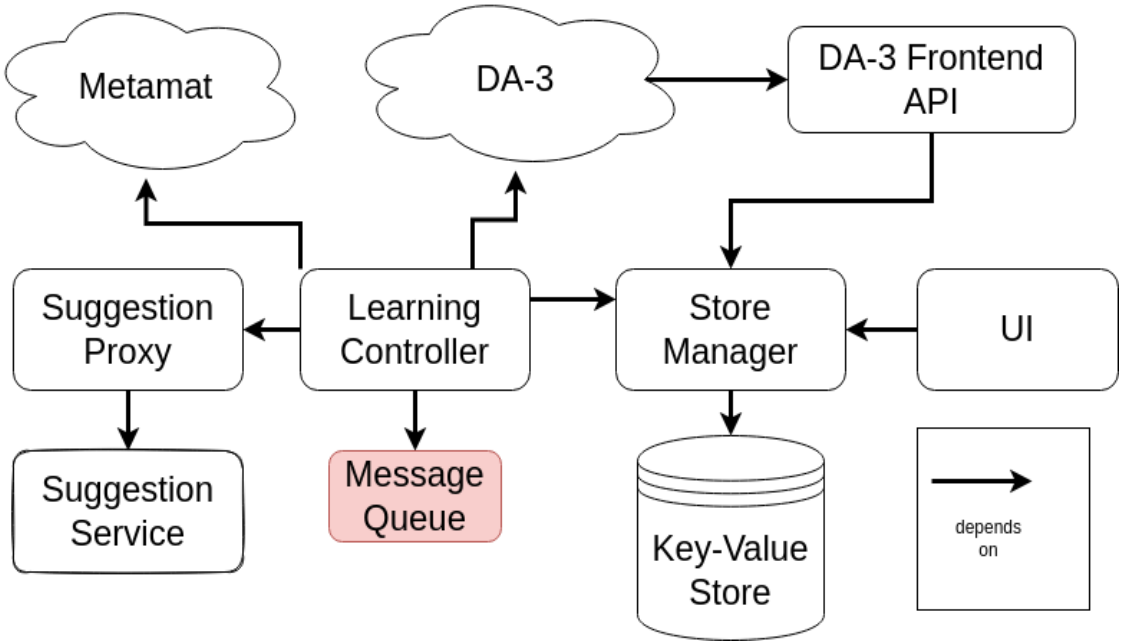
- REST API
- Schema-free
- Precomputed views for queries
- python-cloudant for access



https://couchdb.apache.org
https://github.com/cloudant/python-cloudant

RabbitMQ

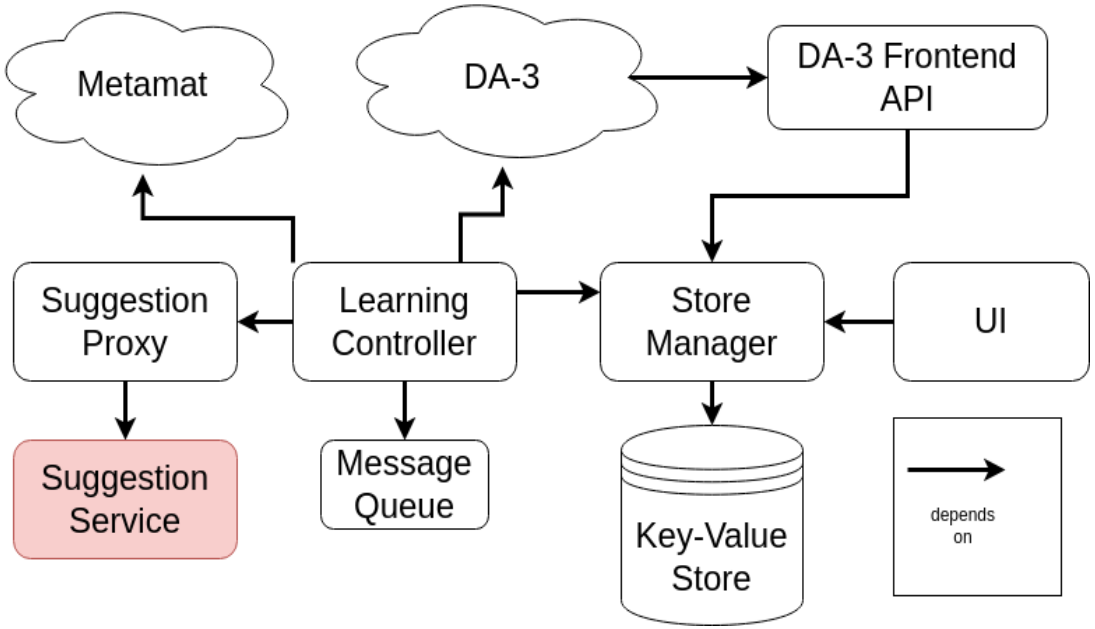
- Easy to deploy & use
- Very popular
- Supports multiple protocols
- pika client library



https://www.rabbitmq.com
https://github.com/pika/pika

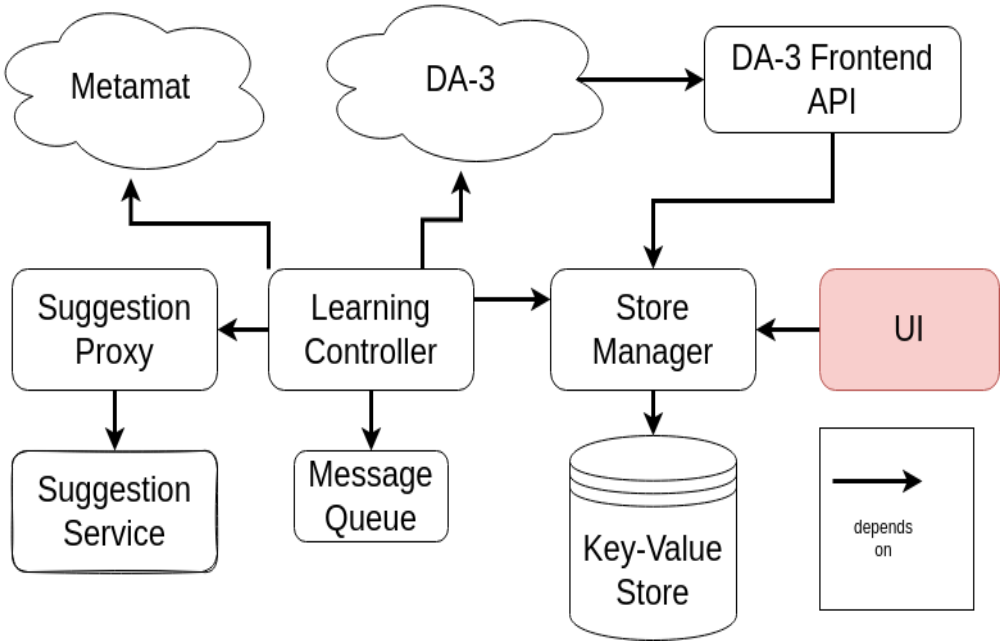
annif

- Toolkit for automated subject indexing
- Used to train our models
- REST API for suggestions



Svelte

- Reactive Web Framework
- Reduced amount of code to write
- Compiles code



Infrastructure

Requirements

- (Almost) No downtime
- Administration by our own
- Inhouse IT only responsible for basic server functionality
- Easy & fast deployments

Cluster

- Kubernetes Cluster
  - 4 Worker Nodes
    - 24 GB RAM
    - 4 CPU
  - 1 Master node
    - 12 GB Ram
    - 2 CPU
- Separation of productive and test system
  - Namespaces
  - Network policies

Storage

- Ceph cluster
  - 600 GB
  - CephFS with 3x replication

- Rook
  - Deployment of Ceph inside Kubernetes cluster
  - Provides Storage Class
  - Supports ReadWriteMany

- Storage Allocation per PersistentVolumeClaim

https://ceph.io
https://rook.github.io

Deployment

- One Kubernetes Pod per Component
- Deployment to test / production directly from CI
- Helm Charts for our Software Architecture
    - Deploy/update all at once
    - Rollback

Monitoring

- Prometheus for alerts and long-term monitoring
- Grafana for visualization
- Elasticsearch – Stack for Logging
- Various cronjobs

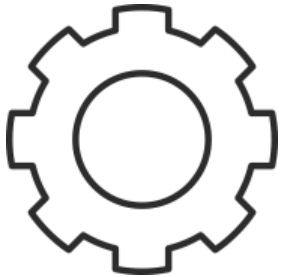
GitLab

- Version Control System
- Collaboration
- Continuous Integration / Deployment
- Container Registry
- Kubernetes Integration

https://gitlab.com

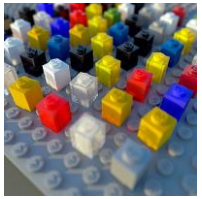
Future Plans

- Canary Deployments
- Publication of UI
- Automation of ML procedures
- Human in the loop



Summary

- Microservices based architecture



- Focus on
  - lightweight technologies
  - easy administration
  - fast deployment

- Container-based infrastructure

Thank you

More information about AutoSE:

https://www.zbw.eu/en/about-us/key-activities/automated-subject-indexing

Contact: {a.kasprzik,m.fuerneisen,c.bartz,autose}@zbw.eu