KRATT-Automated cataloguing National Library of Estonia

(2024/2025)

Urmas Sinisalu
Head of Library Services Center
30.01.2025

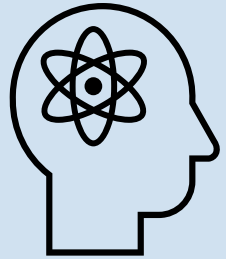
RaRa EESTI RAHVUS-RAAMATUKOGU

UNISTUSTE TÖÖANDJA 2020 OSALEJA

PEREsõbralik TÖÖANDJA

About project

- Project name: **KRATT – KATA Automaatne kataloogimine**
- Project lenght: **12** month
- Total cost with taxes: **596 706€** EU/SF money
  +36 096€ NLE money
- Annual cost in future: **~44 000€** NLE money
- Timeline: 30. July 2024 - August 2025



- Based on the legal deposit files.
- The automatic creation of bibliographic record
  (Title, author, ISBN, page, UDC, keywords)
- helps with digitization, describing web files, etc.

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

Planned to use for all Estonian libraries, memory institutions, ...

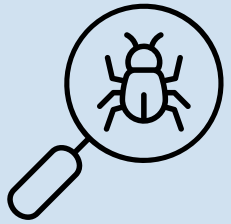
Automated cataloguing

Fully automatic protess for legal deposit files. Modular solution, based on microservices.

- File analyzer to take out text from pdf, html, epub, METS/ALTO
- OCR quality control (+Tesseracti support)
- Detecting metadata (title, authors, year, UDC, ISN, imprint, illustrations etc.)
- Create work, instance records and linking together with authority records (persons, organizations, places)
- Add subject headings
- Find table of content and summary
- Output can be libray system, MARCXML, DC, JSON, CSV, improved METS/ALTO files

Possible to define different workflows (e.g. File->OCR->new record->e-catalog)

Data fields and their identification rules



General Information:
•Language of the publication
•Authors of the publication
•Title of the work

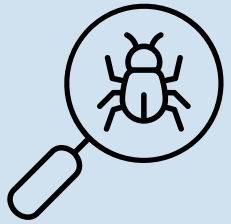
Identifiers:
•ISBN
•ISSN
•URI
•Edition

Relations:
•Work to Edition
•Edition to Work

Publication Information:
•Publisher, Place of publication, Year of publication
•Printing house, Place of printing, Year of printing
•Distributor, Place of distribution, Year of distribution
•Copyright year

Physical characteristics:
•Dimensions, number of pages, Illustrations
•Content of the publication
•Medium
•Carrier
•File format, structure, size, resolution

Data fields and their identification rules



Content Information:
- Table of contents
- Brief description
- Edition
- UDK (UDC)

Series Information:
- Series information

Keywords:

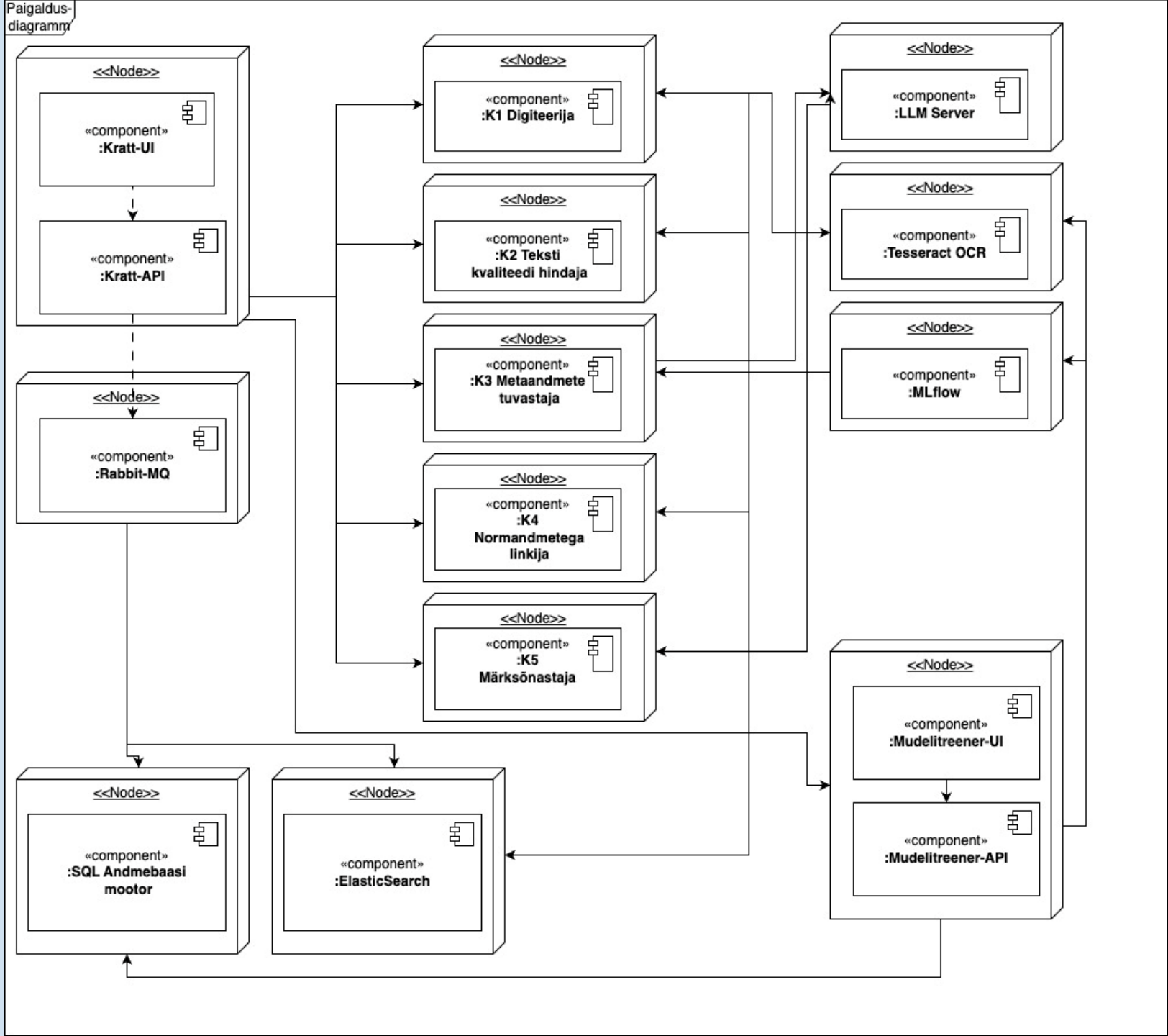
- Individuals and works as keywords
- Collectives as keywords
- Subject keywords (Content keywords)
- Form keywords (genre keywords)
- Time keywords
- Place keywords
- Temporary collective (event as a keyword)
- Authorless/anonymous publication as a keyword
- Other keywords

Technical solution

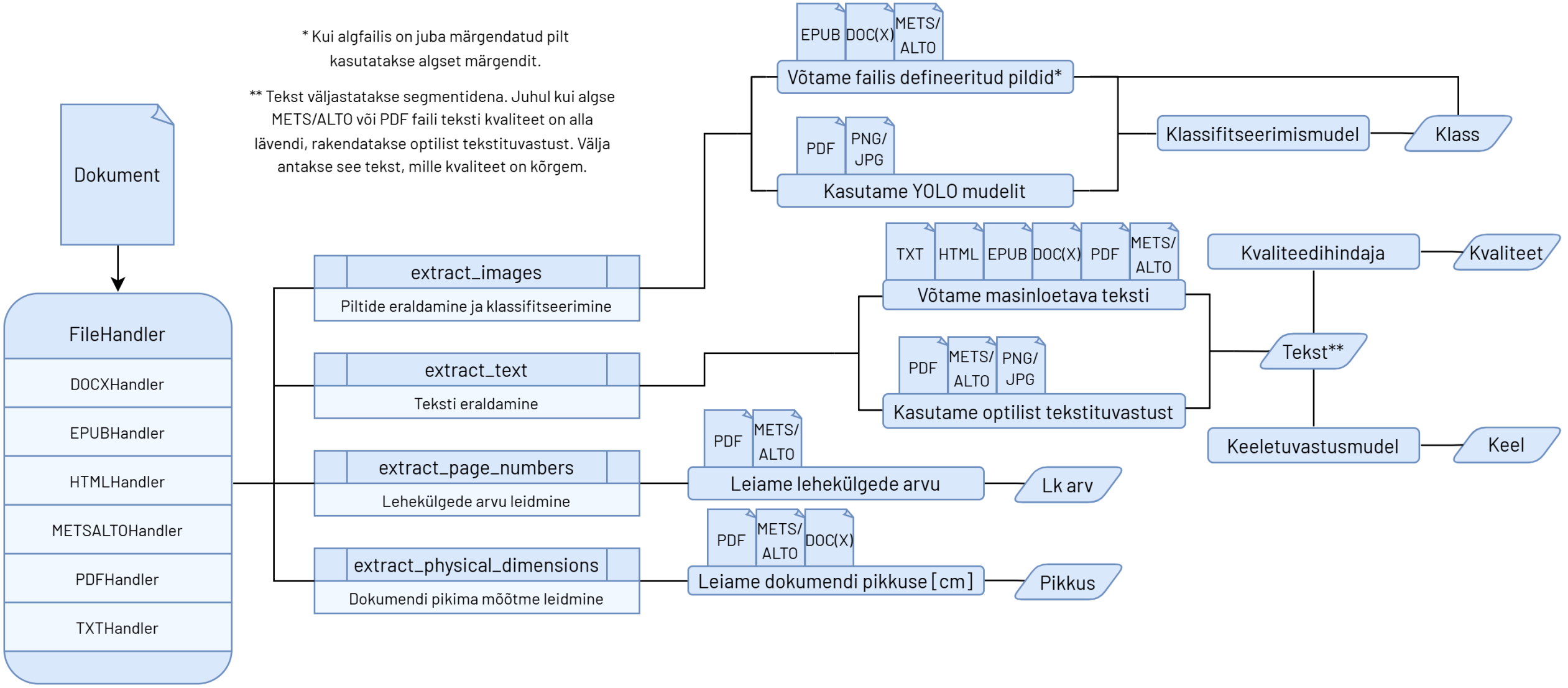
- CORE module (UI+API+Priority handler)

- K1 – File handler
- K2 – Text quality evaluator
- K3 – Metadata extractor
- K4 – Data linker
- K5 – Subject indexing

- Model trainers



K1 - File handler



Text quality evaluator

Supported languages:

- Estonian
- English

Detecting most of languages and mapping to MARC21

Ex:

"languages": [
{"language": "et", "count": 344, "ratio": 0.994},
{"language": "en", "count": 1, "ratio": 0.003},
{"language": "de", "count": 1, "ratio": 0.003}
],

Tekst	ET	EN	X
• l * QMtTCi üjSSjS[§B§B1 laErljfctey I 1|L .. WWMW ÄTWRI I* * flOSt R;5!iWB®? jIt fl MMfc W '	0.13	0.11	0.12
TÄNA KUKUS! TALLINNAS 100,7 MHz RAADIO JÄRVAMAAL 100,5 MHz MÕTLEVALE MULGIMAAL 100,8 MHz INIMESELE TARTUMAAL 101,2 MHz jfVjk /c 17-11L30 I http://www.zzz.ee/kuku/ |	0.7	0.22	0.66
Päewaleht 31mub Igal arlpSewa!. Tallinnas, Suurel turul, raekoja ees, Punkri uulltfd nurgal, nr. 4. Talmctufc Mflttr&if W Kk?. Toimetaja bduetund fertla 13—& JMrefs liht- Ja rahaklrjadelet ?Sewalehe toimetutele, Tallinnas, erb r. Pcne/ib. Kontori Mnetraat «I 390. ftrtjuhl harutuna hella S —l 2.	0.48	0.14	0.39
Tänavu aasta on katk Astrahani kubermangus mitmel pool avalikuks tulnud. Pea-arsti-valitsus sai professor Sabolotni käest teate, et kõik, kes katku kätte haigeks jäid, ka on ära furnud. Terveks ei ole ükski saanud. Professor	0.94	0.21	0.86
If you only walk on sunny days, you will never reach your destination. There is no elevator to success, you have to take the stairs.	0.81	0.96	0.98
Yos, mnn os martal, but thot isx’t sa bad. Wkat's bad is that soxretimes ke's unexpeetwdly mortal, thats the rut. Aud, in general, he cant eveb say io the moming what hell he doiug that vwy sawe nighf.	0.49	0.44	0.68
Tekst lühem kui 30 tähemärki.	0	0	0

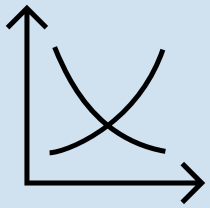
Good vs Bad

Must be completly closed system ☹

- No precise definitions for **Material Type** (book vs. standard vs. ephemera)
- No good **local** authority systems or files (authors, places, corporate bodys)
- Can’t use big LLM for **legal deposit** files.

- Good **Estonian Subject Thesaurus** (https://ems.elnet.ee/)
- Good **models** for detecting illustration types (graphs, maps, photos etc.)

Detecting illustration types



TOP 7

- Stamps
- Graph/chart
- Maps
- Photo
- Etiquette/label
- Ex libris/bookplate
- Caricature

Classification Report:

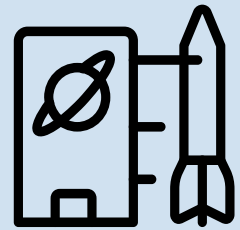
	precision	recall	f1-score	support
akvarell	0.62	0.80	0.70	199
eksliibris	0.88	0.92	0.90	199
etikett	0.94	0.86	0.90	199
foto	0.93	0.87	0.90	199
graafik	0.97	0.96	0.97	199
graafika	0.82	0.52	0.63	199
joonis	0.82	0.89	0.85	199
joonistus	0.59	0.78	0.67	199
karikatuur	0.93	0.87	0.90	199
maakaart	0.98	0.88	0.93	199
maal	0.83	0.67	0.74	199
mitteillustratsioon	0.86	0.86	0.86	199
pitsat	0.99	0.95	0.97	199
plakat	0.81	0.86	0.84	199
postkaart	0.76	0.81	0.78	199
varia	0.84	0.83	0.84	199
accuracy			0.83	3184
macro avg	0.85	0.83	0.84	3184
weighted avg	0.85	0.83	0.84	3184

Still in progress ☺ …

- All modules are still in analyze stadium and not ready jet. Hopefully I can speak as workin solution at the end of this year and then I can go more deeply in technical details.

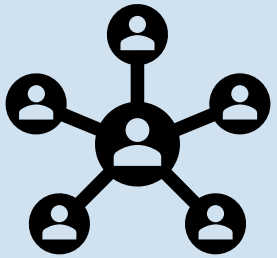
KRATT - Future plans

(2025/2026)



In cooperation with the National Library of Finland/University of Helsinki

- Adding an alternative keywording module (Annif)
- Building webarchive description capability
- Extend input file types and output formats (.tsv)
- Add API support for DSpace
- Include other memory institutions to find more use cases (improve knowledge base and training data)



Let's explore together!

digilab.rara.ee

Thank you and more questions please!

Urmas.Sinisalu@rara.ee