WEBCRAWLER FOR HYDROLOGICAL DATA

ECMWF Summer of Weather Code

Norwin Roosen @ 52° North
ESoWC Webinars, 20.09.2018

THE PROJECT

Web Crawler for hydrological data

- goal: “*Development of a tool to search the web systematically, identifying data sources for observed environmental data*”

solution:

- multilingual search engine, but with thematic focus
- classification of web pages (contains/links to data?)
- data set metadata extraction

CHALLENGES: UNSTRUCTURED DATA

- only some data on the web is listed in catalogs

  → easily accessible, not our focus

- data without machine readable annotation hard to discover
  - links to data without proper annotation
  - embedded into HTML pages
  - access on request only
  - web (map) applications → undocumented APIs

CHALLENGES: MULTILINGUAL SEARCH

- global data coverage is helpful for predictive models
  - effective search in local languages only

→ need to reduce language barriers via translation of

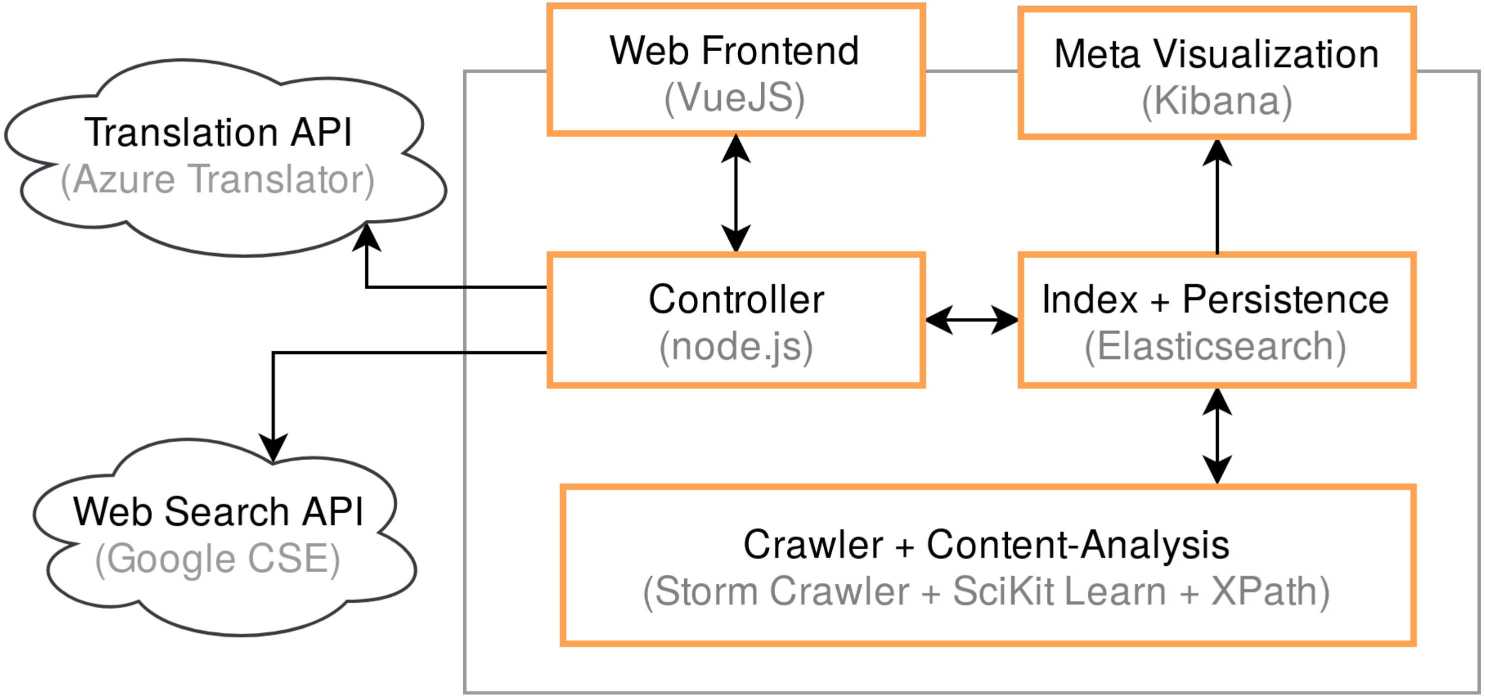
  - search keywords
  - page content

→ language independent content analysis

FEATURES

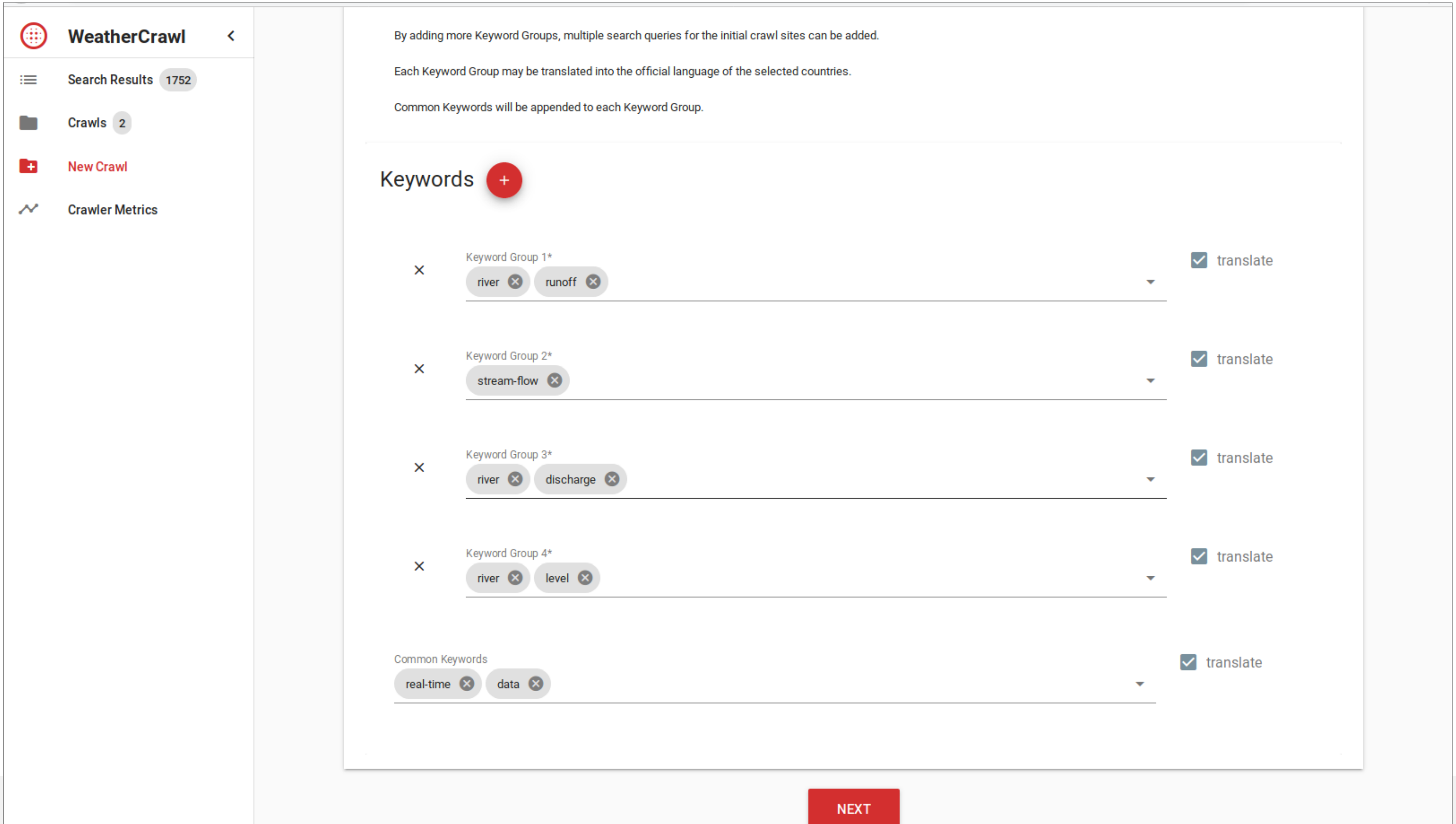
- web search
  - automated translation of search terms by country
  - crawling of outlinks for broader spectrum
- content analysis
  - classification via custom machine learning model
  - metadata extraction
- result list & crawl management via web-UI
  - result list with extracted metadata & hotlinks to translated pages
  - result labeling interface for classifier training
  - performance metric visualization

ARCHITECTURE



- job based: minimize crawl effort by crawling on specific topics only
- stream based: results accessible as they come in

DEMO



By adding more Keyword Groups, multiple search queries for the initial crawl sites can be added.

Each Keyword Group may be translated into the official language of the selected countries.

Common Keywords will be appended to each Keyword Group.

Keywords +

Keyword Group 1*: river, runoff ☑ translate

Keyword Group 2*: stream-flow ☑ translate

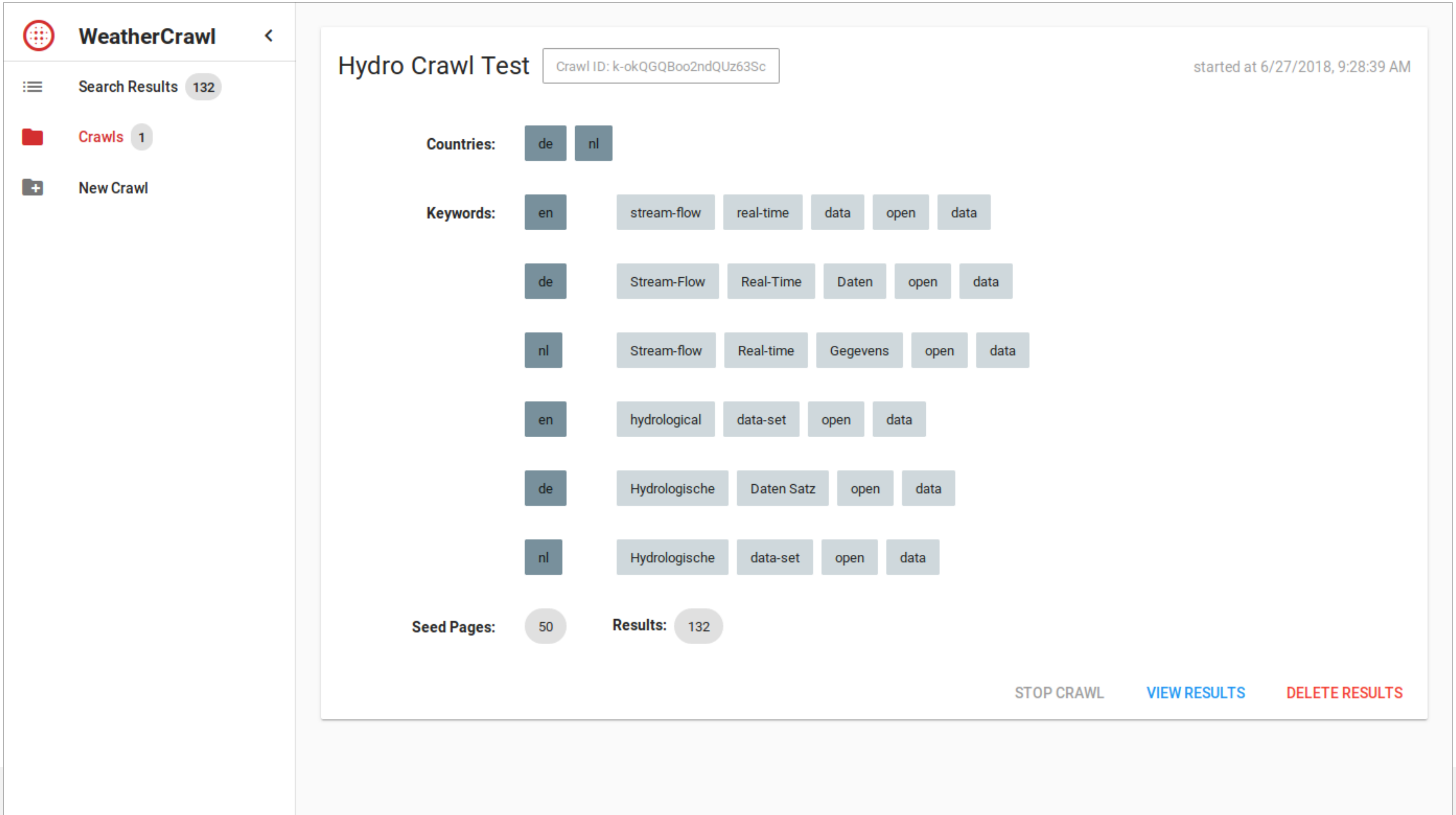
Keyword Group 3*: river, discharge ☑ translate

Keyword Group 4*: river, level ☑ translate

Common Keywords: real-time, data ☑ translate

NEXT

DEMO



Hydro Crawl Test

Crawl ID: k-okQGQBoo2ndQUz63Sc

started at 6/27/2018, 9:28:39 AM

Countries: de nl

Keywords:

en stream-flow real-time data open data

de Stream-Flow Real-Time Daten open data

nl Stream-flow Real-time Gegevens open data

en hydrological data-set open data

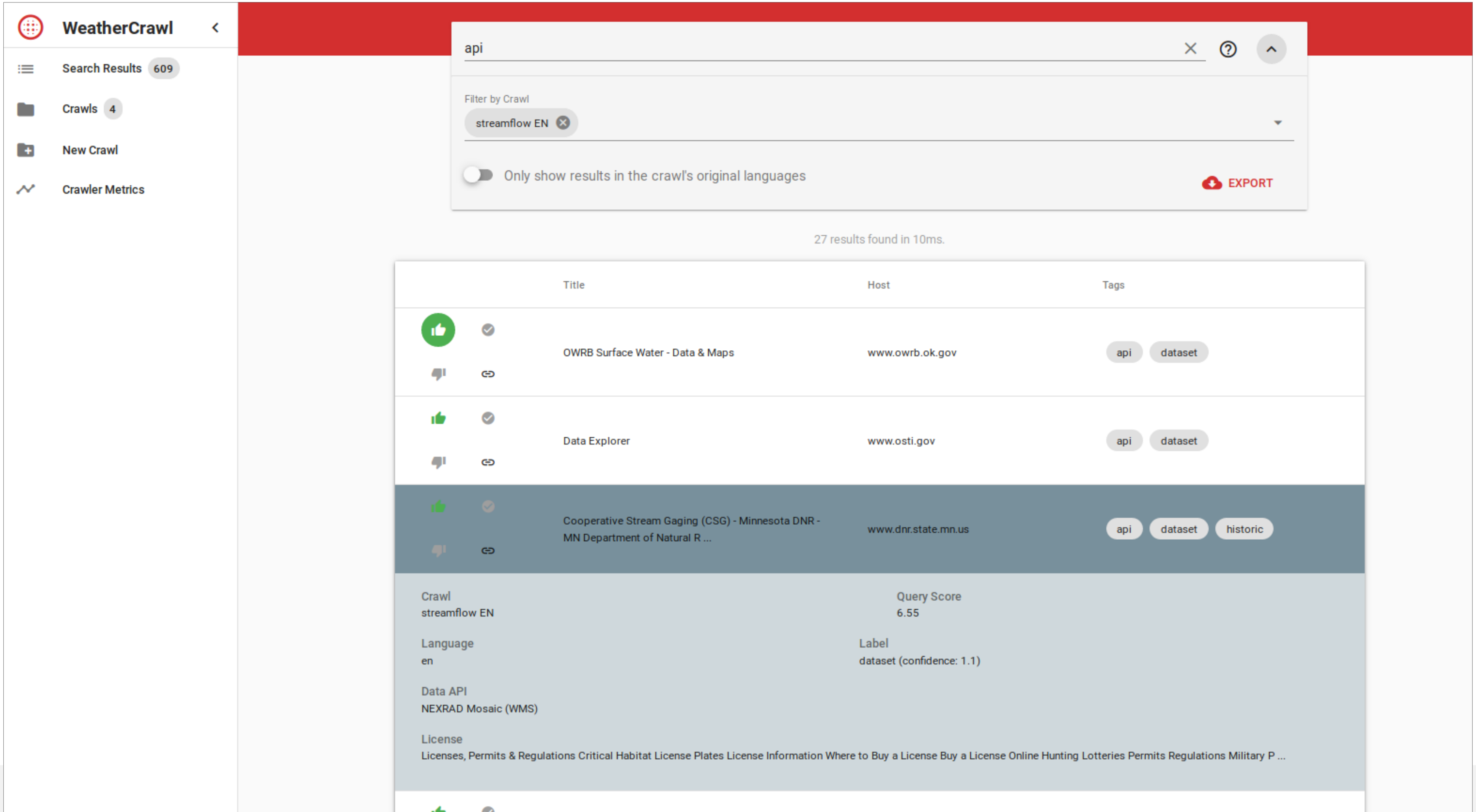
de Hydrologische Daten Satz open data

nl Hydrologische data-set open data

Seed Pages: 50

Results: 132

DEMO



api

Filter by Crawl

streamflow EN

Only show results in the crawl's original languages

27 results found in 10ms.

Title	Host	Tags
OWRB Surface Water - Data & Maps	www.owrb.ok.gov	api dataset
Data Explorer	www.osti.gov	api dataset
Cooperative Stream Gaging (CSG) - Minnesota DNR - MN Department of Natural R ...	www.dnr.state.mn.us	api dataset historic

Crawl
streamflow EN

Query Score
6.55

Language
en

Label
dataset (confidence: 1.1)

Data API
NEXRAD Mosaic (WMS)

License
Licenses, Permits & Regulations Critical Habitat License Plates License Information Where to Buy a License Buy a License Online Hunting Lotteries Permits Regulations Military P ...

REVIEW

Were the initial goals met?

✔ effective discovery of websites & data sets

✔ language-abstracting interface

✗ classification not operational for every language

REVIEW

Future work

- train classifier for more languages
- check more dataset annotations (→ schema.org)
- performance optimisation of crawler

THANKS!

- Code: Apache 2 licensed on GitHub
  github.com/52north/ecmwf-dataset-crawler
- Read more in the..
  - project wiki: github.com/52north/ecmwf-dataset-crawler/wiki
  - 52N blog: blog.52north.org/2018/06/26/ecmwf/