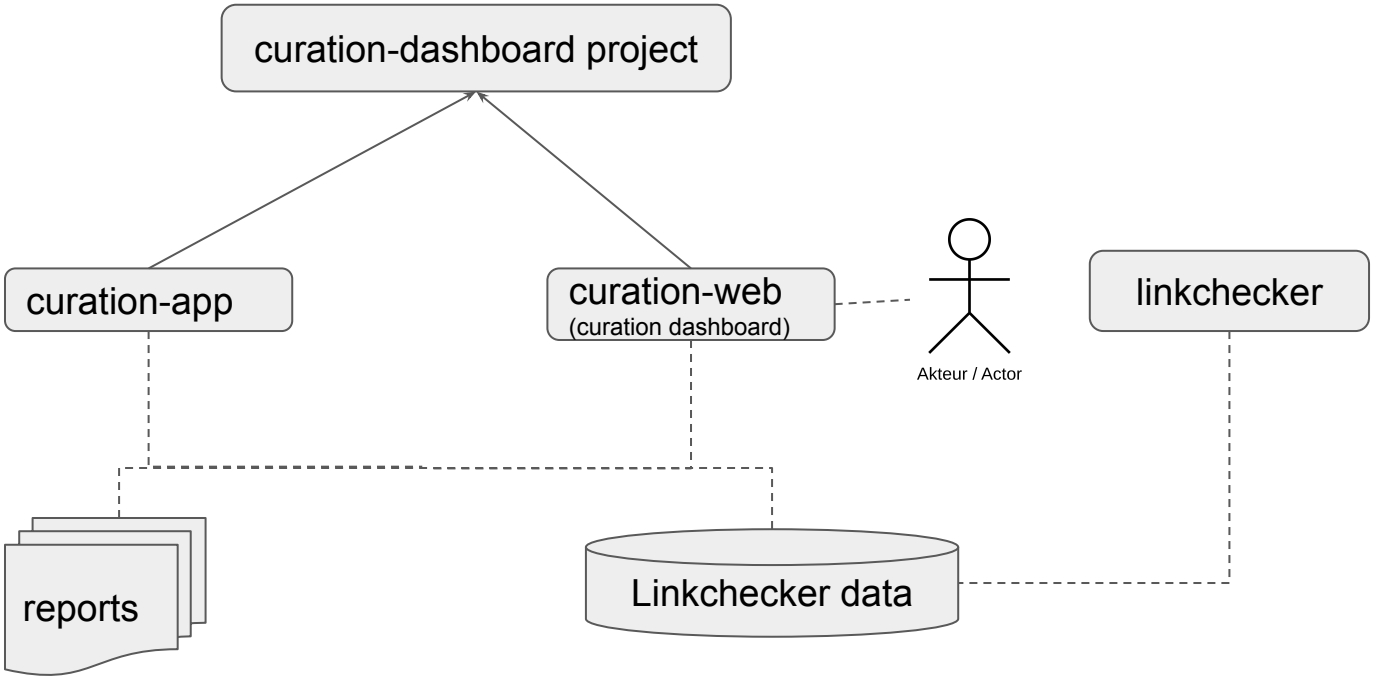
curation dashboard and linkchecker

what it is and how to use it

these notes:
https://docs.google.com/presentation/d/13HCDNyUf-N9neAPHWsYxAwRDadoKvzTAT2dlJ9iT190/edit?usp=sharing

Overview


curation-dashboard project
curation-app
curation-web
(curation dashboard)
Akteur / Actor
linkchecker
reports
Linkchecker data

curation dashboard

production: https://curation.clarin.eu

development: https://alpha-curation.clarin.eu

Statistics (state 2024-06-10)

- reports generation 4-times per week
- 48 collections
- 746,951 files
- 3,447,924 links (2,977,186 unique URLs) from resource or md-selflink

Validator

- upload of profile or CMD instance from your computer or via URL
- for information on scoring see https://curation.clarin.eu/faq

Profiles

- generates reports from all published cmd 1.2 profiles with status production, development or deprecated
  (https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles?registrySpace=published&status=production&status=development&&status=deprecated)
- has usage information

Collections

- aggregates single CMD file information to a collection result
  - slight difference: link checking results and score only available for collection
    => maximum score for individual CMD file is 14, while 15 in a collection
- identifies files with issues

Link checking

- presents link checking results per collection
- groups results in 6 categories:
  - Ok (200, 304)
  - Undetermined (405, 429)
  - Restricted access (401, 403)
  - Blocked by robots.txt
  - Broken
  - Invalid URL

Some facts on link checker application

- stand alone application, based on apache storm and storm crawler
- runs permanently with 50 threads, 1 thread per host
- default crawl delay of 1 second per host
- strictly respects robots.txt
- re-checks are blocked for 24 hours
- checks currently about 200k links per day on about 800 hosts