Processing Crawled Data

OSF's OSSYM Oct 11, virtual
Mark Overmeer, Skrodon

Road Trips

- End of june 2021:
  - Michael Granitzer
  - Iosif Peterfi
  - Stefan Voigt
  - Sebastian Nagel
- September 2021:
  - Bert Niehaus
  - Stephan Schwichtenberg
- Virtual:
  - Djoerd Hiemstra
  - Olivier Blanchard

Conclusions

- Really, really nice and enthusiastic people!
  Thank you for your hospitality!

- Willing to contribute, but where and how?
  - Fits with research obligation?
  - Man-power?
  - Continuation?
    → production, maintenance, support

Meanwhile,

- Team Skrodon +=
  - Красимир Беров
  - Ronny Lam

- Cooperative projects
  1) Crawl Pipeline → now
  2) Who Has What
  3) Crawl Planner
  4) Open Console → Tue

- *Open* Source & Infra

„Skrodon”

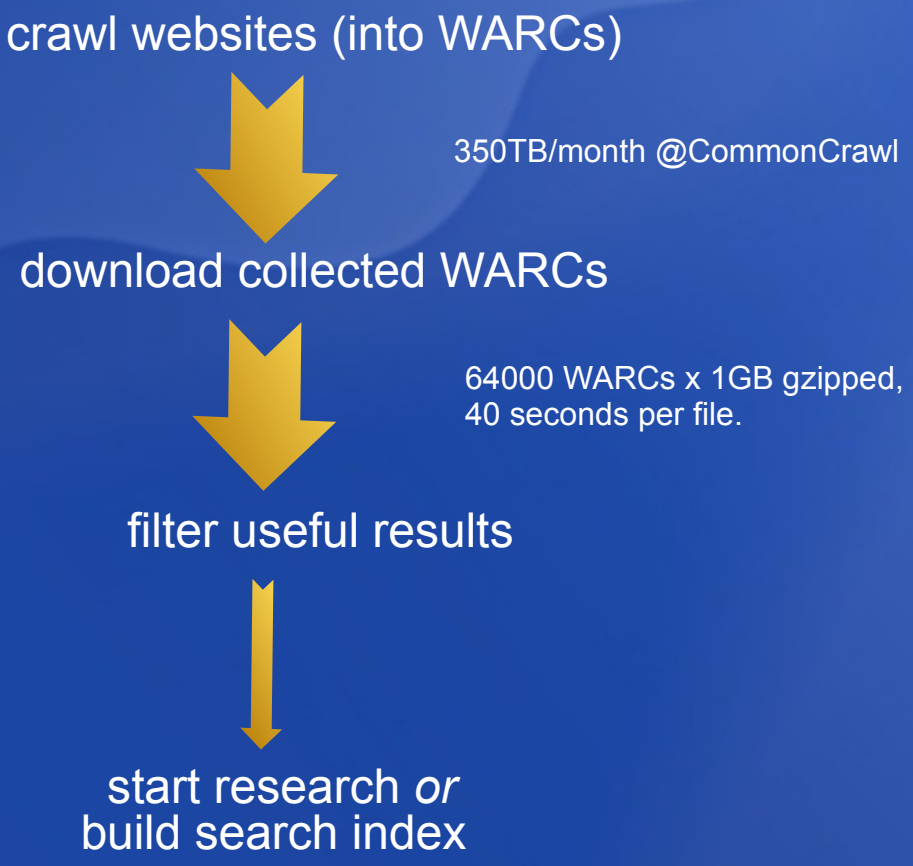
- Share real website related data, on full internet scale. Open Infrastructure
  - Share crawled collections
  - Share extracted (meta)data
  - Share computed data / computation resources

  → research can focus on extract

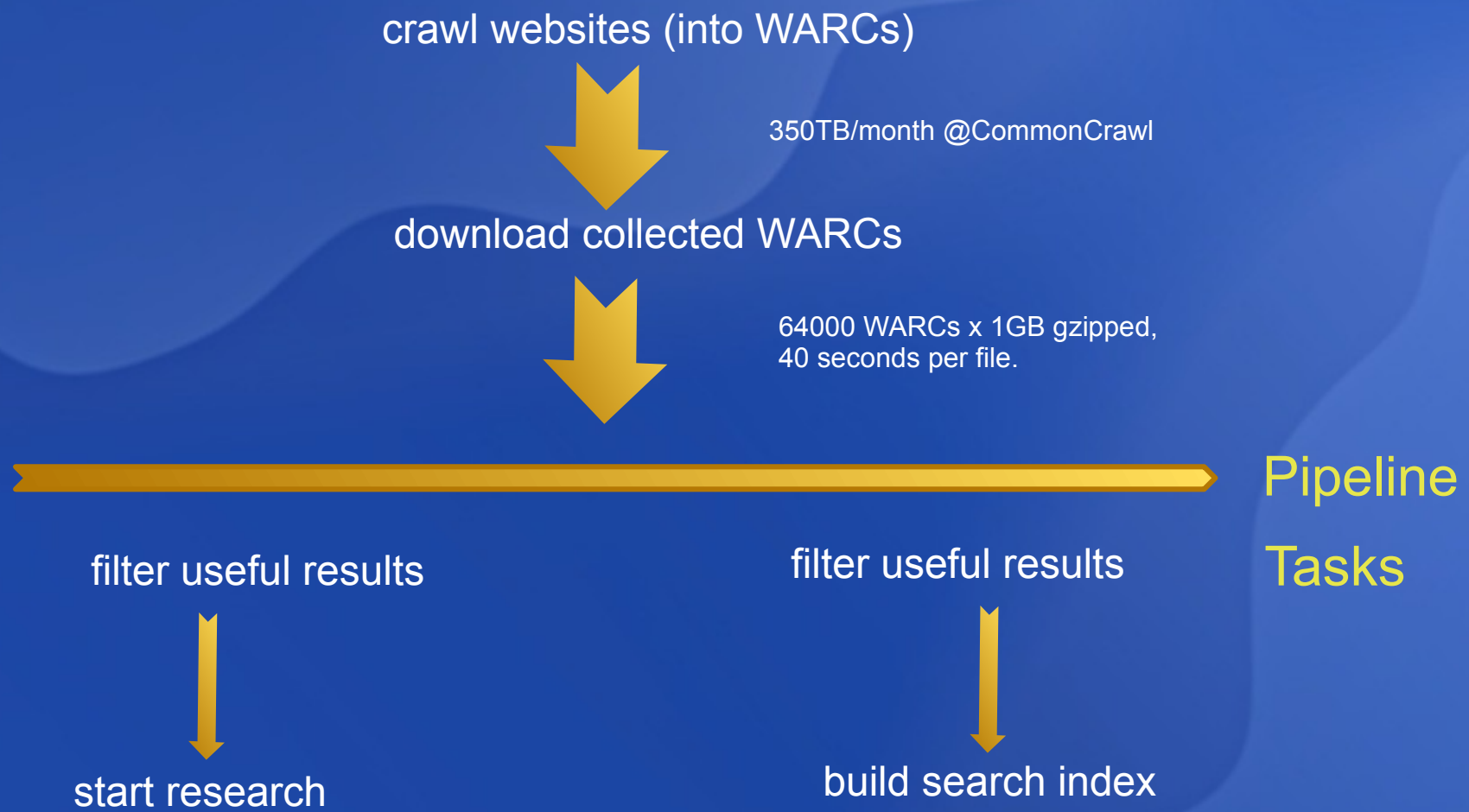
„Skrodon”

- Share real website related data, on full internet scale. Open Infrastructure
  - Share crawled collections
  - Share extracted (meta)data
  - Share computed data / computation resources
  - Strict EU law & jurisdiction

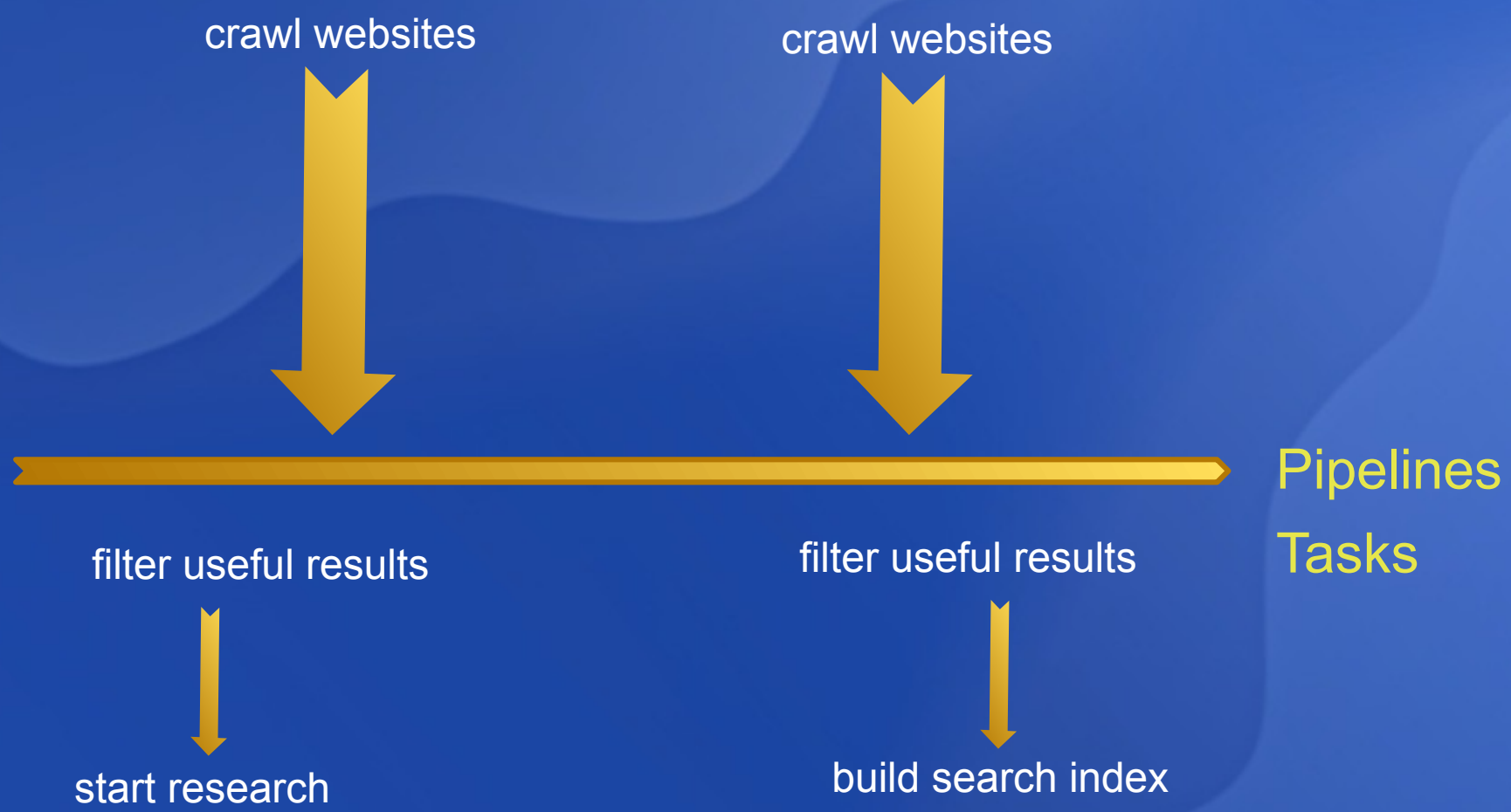
1) Crawl Pipeline



Crawl Pipeline



Crawl Pipeline



Tasks

- Selection Filter
- Data Extraction
- Result Distribution

Tasks; selection

- Selection Filters:
  - language
  - pattern in URL
  - response Content-Type
  - size of text
  - words / patterns in text
  - …

Tasks; extraction

- Data Extraction:
  - request, response, metadata (plain or WARC)
  - CommonCrawl meta (WAT) or texts (WET)
  - HTML inspection
    - <meta> <link> hrefs errors
    - OpenGraph
  - …
- Targeted extraction of the above.

Tasks; distribution

- Result distribution:
  - serialize as XML, JSON, YAML
  - wrap as WARC-record or separate files
  - package as tar, zip, WARC-archive
  - publish via http, https, ftp, src, rsync

  - download from every pipeline, at least every day.

Example

- Filter all webpages
  - in Bavaria
  - written in German
  - about food
  - with at least 300 words
  - which are „quality”
- Extract
  - plain text
  - why selected
- Distribute
  - file per result
  - packaged as zips <1GB
  - published via ftp

Example

- Filter all webpages
  - in Bavaria
  - written in German
  - about food
  - with at least 300 words
  - which are „quality”
- Extract
  - plain text
  - why selected
- Distribute
  - file per result
  - packaged as zips <1GB
  - published via ftp

Project Open Console
Talk „understanding websites”
Tuesday 11:00

More examples

- Count types of CMS used
- Count types of http servers
- Implement OpenGraph search
- Collect links to be crawled
- Report link errors to site owners
- Extract word-use per language

More examples

- Detect languages offered by a website
- Auto-detect phishing attempt
- Auto-detect SEO spam network
- Auto-detect porn/racism/gambling/... sites
- Build auto-completion tables
- ...

Status

- Pipeline for CommonCrawl can run on my PC at home (6 cores) without HTML inspection. <100GB disk, network 55MB/s.
- With full HTML extract requires 17 cores.
- Waiting for your tasks!