Crawler

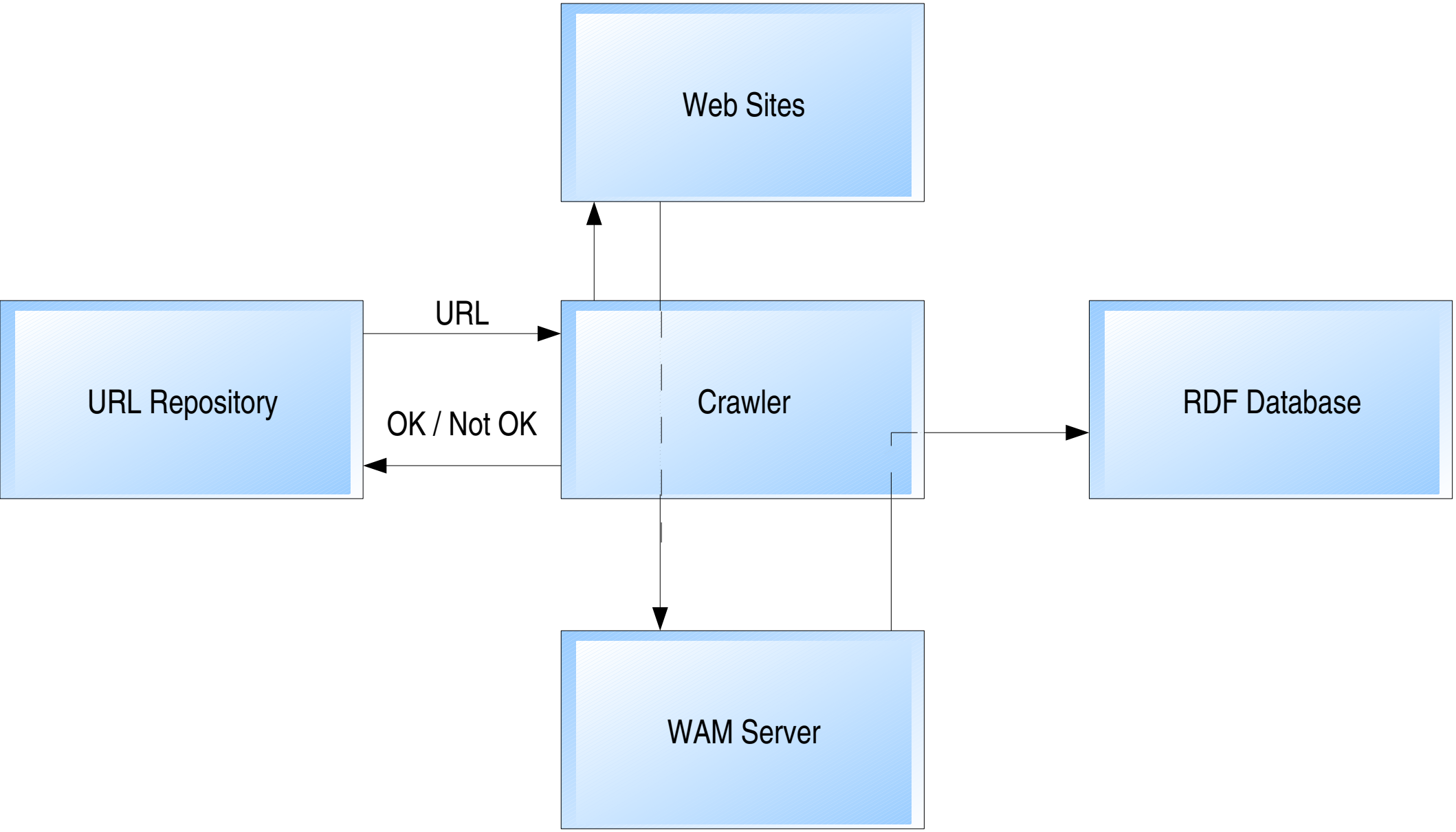
Crawler

Morten Goodwin Olsen

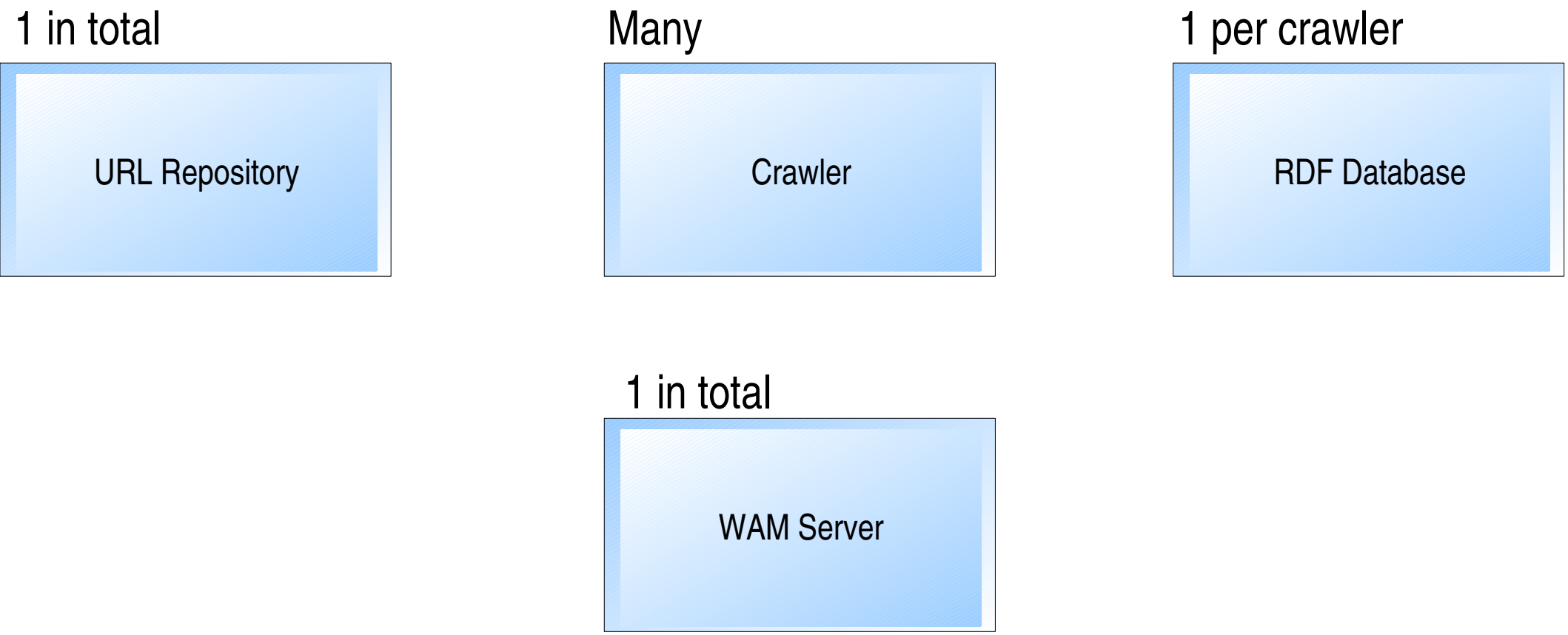
Overview

- Functionality overview
- Possible scalability issues
- Testing
- Open Source Release
- WCAG2.0 / UWEM2.0 migration

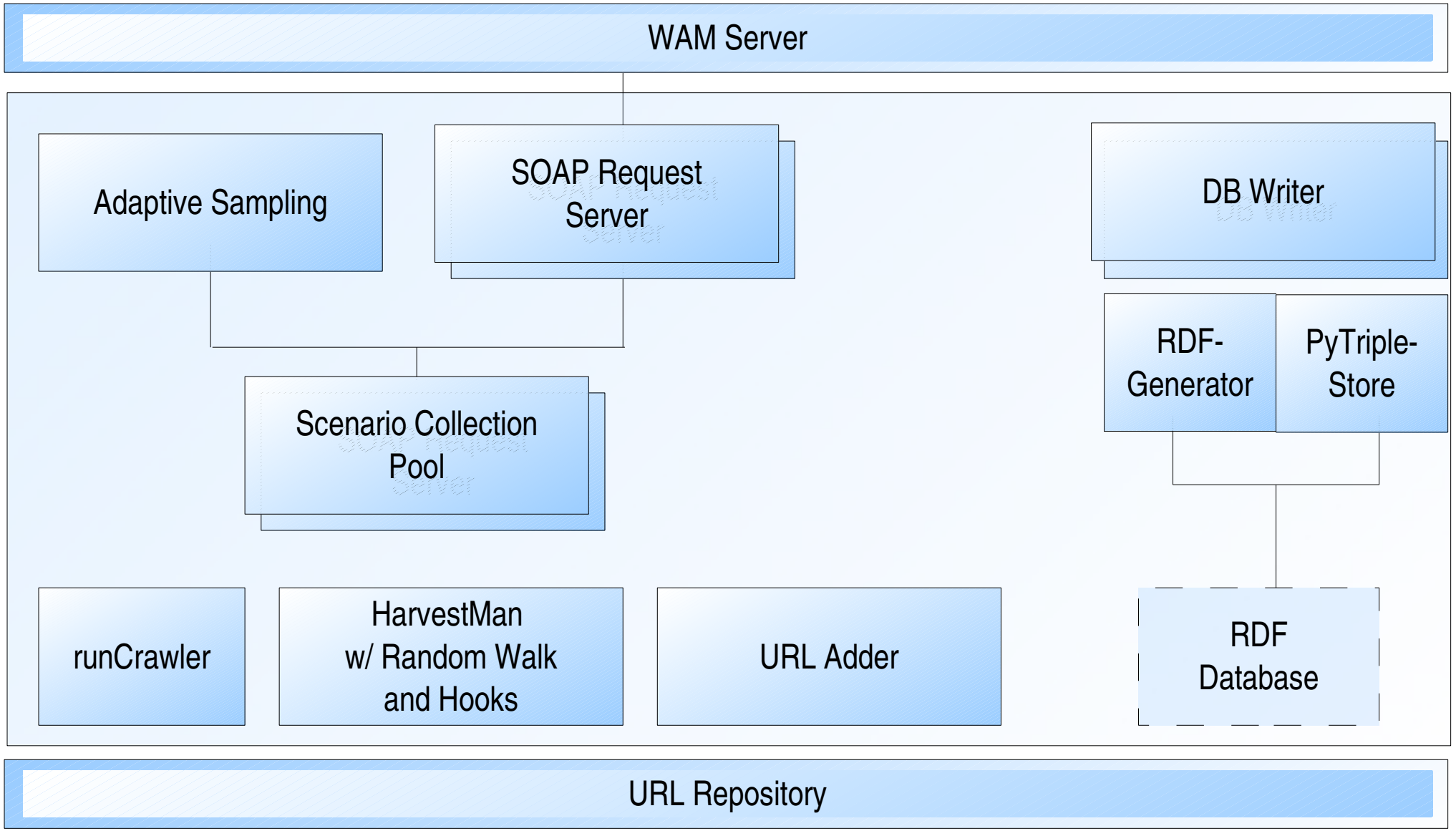
Functionality Overview



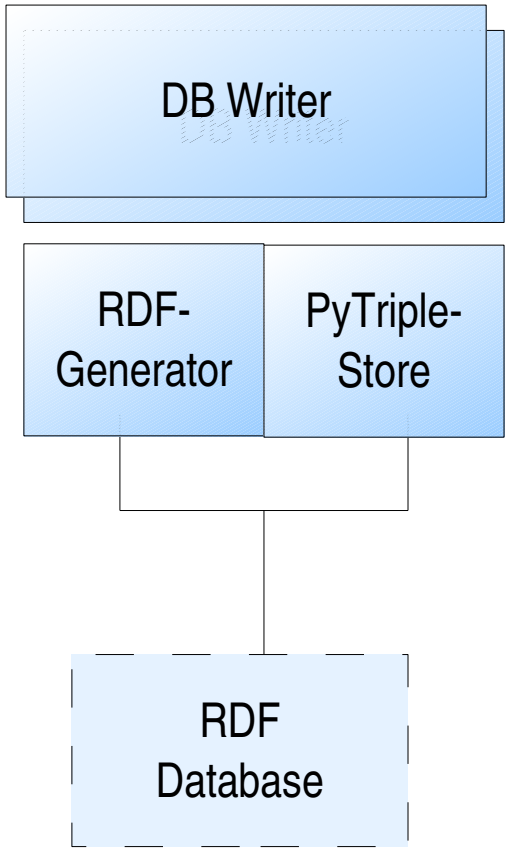
Functionality Overview



Functionality Overview

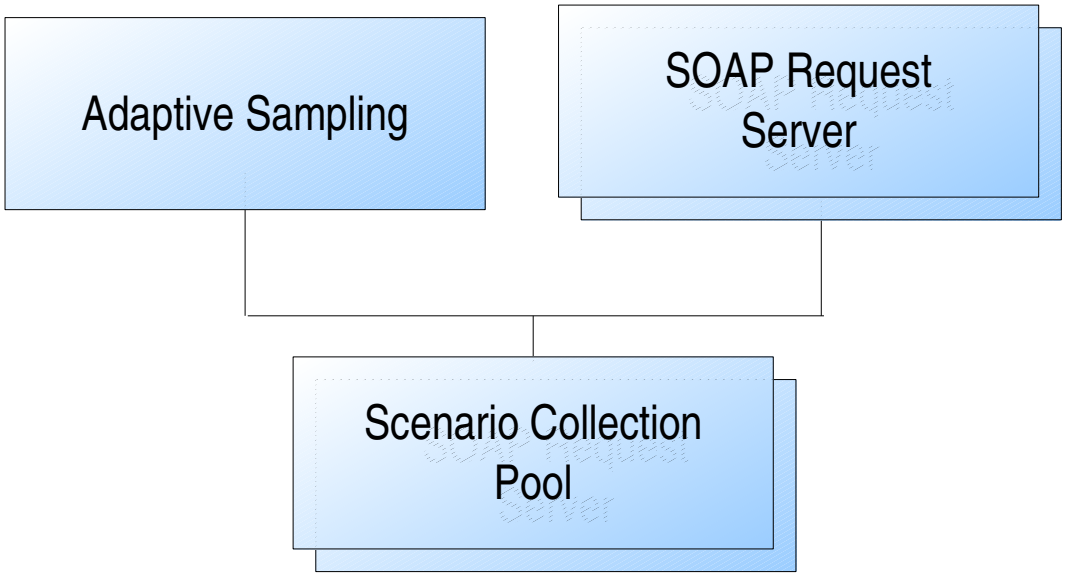


Possible Scalability Issues



- RDF
-
- Writing speed too slow;
  - Should not be a problem when we have the following;
    - Small enough RDF-repositories
    - Asynchronous writing

Possible Scalability Issues



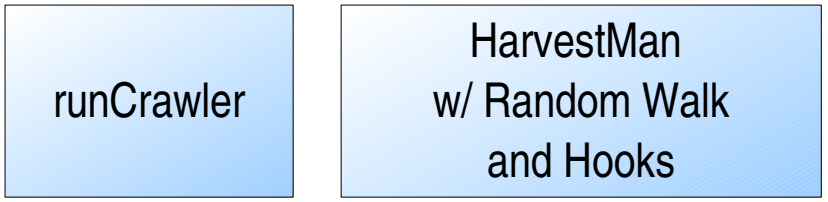
- Interacting with WAM
  - Too much idle time waiting for the WAM?
    - With asynchronous calls through the SOAP Request Server, this should not be a problem

Possible Scalability Issues

Adaptive Sampling

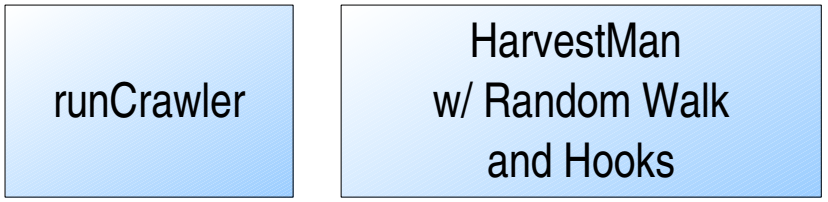
- Oversampling
  - Since we run a multithreaded system, when an error margin is reached, there are always some samples left. This will in most cases reduce the error margin (which is good). However, in some cases the error margin may increase.
    - Possible solution: Sample to an error margin of 0.04 instead of 0.05 to be almost sure.

Possible Scalability Issues



- Some crawlers hang due to bugs we do not yet know
  - Hardly no issue since the hanging crawlers are killed.
  - No individual hanging crawlers will stop the entire crawls, because we are running several crawler processes at the same time.
  - Fix Bugs when they appear

Possible Scalability Issues



- Unknown error 514 (and previously unkown error 512)
  - Seems to be avoided when we force the crawler to restart after each 100th crawl.
    - Just a workaround, not a solution.

Testing

Suggested tests;

- Able to crawl 10 sites (passed)
- Able to crawl 100 sites (passed)
- Able to crawl 1000 sites (passed)
- Able to crawl 10 000 sites

- Result, Error margin, variance, standard deviation correct

- No memory issue
- No storage issue

- Able to “crawl” an unavailable site
- Able to crawl a normal site

Testing

Suggested tests;

- Two crawls towards the same unchanged site gives roughly the same result.
- Calculated result and error margin is the same in the Crawler and the recalculated results in the DW.
- Calculated result is correct even if several scenarios fail.
- Detect that CSS and (X)HTML are gathered correct in the scenarios

Testing

Suggested tests:

- Random walk is random (same as i r1)
- Correct time stamps
- All data is present in the RDF-graph (meta data, WAM meta data, WAM results, Imergo WAM results)

Scalability tests

- Hotshot analysis
- “Timestamp”-test (same as in r1)
- Cyclomatic complexity (?)

Open Bugs

- #407 Unkown error 514... :(
- #372 Error in get_full_url(...).
  - Not seen in a long time, but not fixed yet.
- #365 CSS sampling does not support ignoring any @import rules that occurs inside a block or that doesn't precede all rule sets
- #193 HTTP Get.
  - This is to the best of my knowlede fixed

WCAG 2.0 Migration

- Sampling
  - Will current sampling strategies remain?
  - How to sample from e.g. a PDF, Flash?
    - Entire PDF / Parts of the PDF
- Other technologies (In addition to (X)HTML/CSS).
  - What is a page?
  - Can the current assumptions of a page be