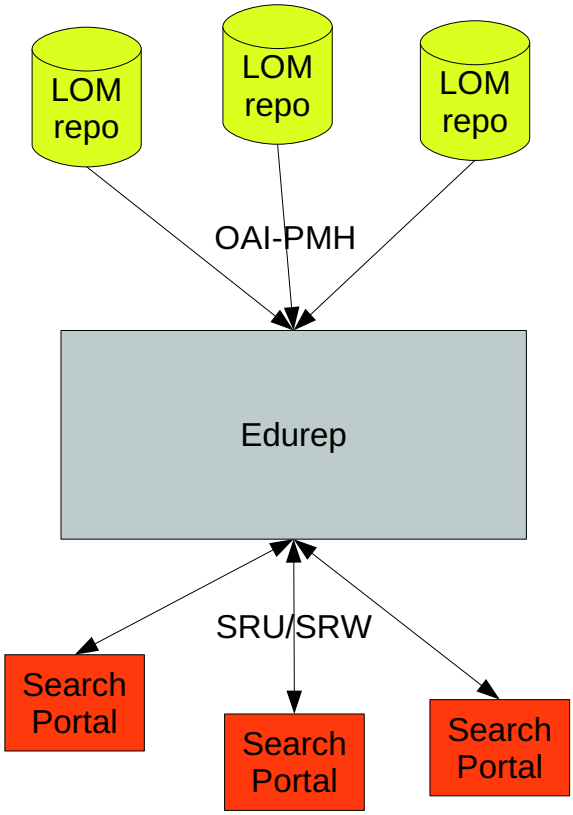
Where is the user? Filtering Bots from the Edurep Query Logs

Wim Muskee

Kennisnet Foundation

SE@M 2010

Edurep in Context



- harvesting more than 50 repositories
- almost 800.000 records
- queried by more than 10 production portals
- more than 500.000 queries each month

  (peak > 2.000.000)

Query Language

- CQL 1.1 over SRU/SRW

- query=elephant
- query=elephant OR mammoth
- query=lom.general.title=elephant
- query=snake AND lom.educational.context=PO
- query=lom.lifecycle.contribute.publisherdate>2009

more query examples

- Pagination:

  query=dog

  query=dog&startRecord=11&maximumRecords=10

  query=dog&startRecord=21&maximumRecords=10

- Specific records:

  query=lom.general.catalogentry=memorix:na:col1:dat148993

  query=meta.upload.id=na:oai:mdms.kenict.org:217136

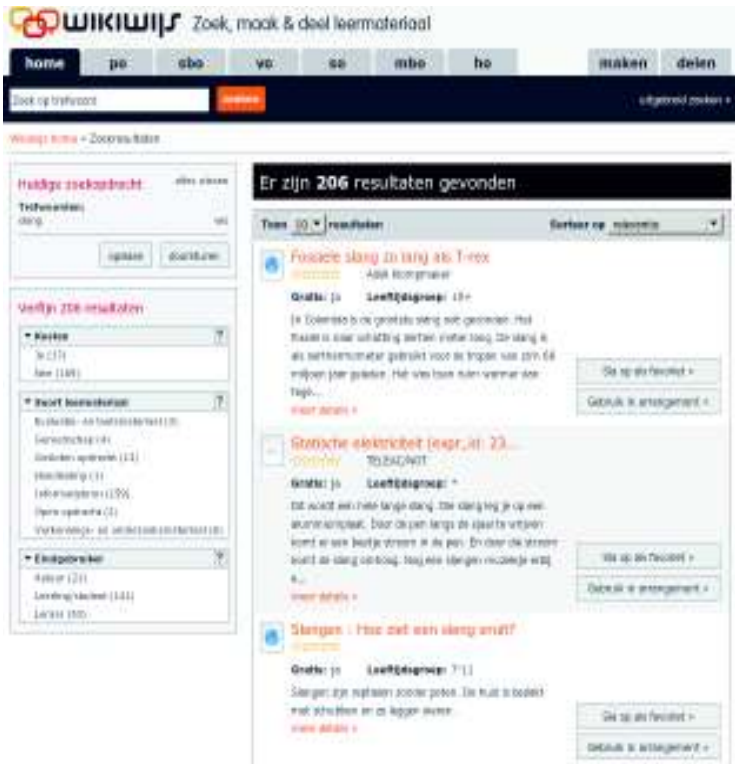
- Facets using facet index:

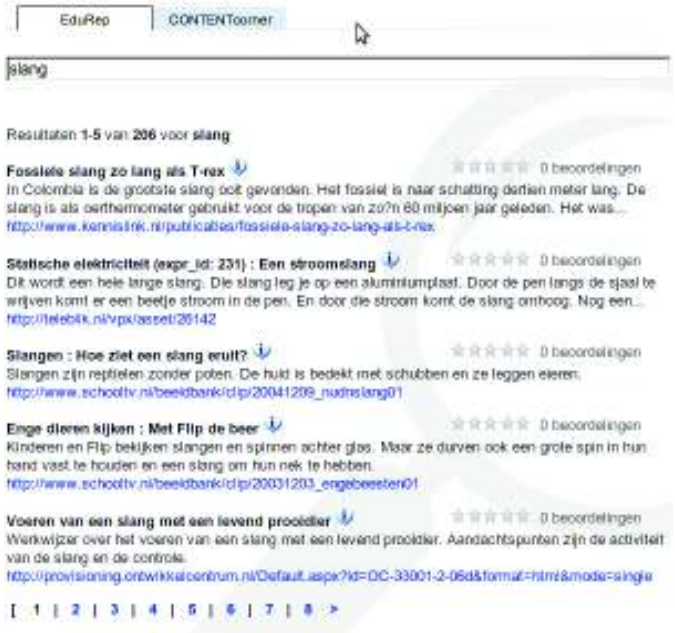
  query=dog&x-term-drilldown=lom.technical.format

XML Query Response


```
- <srw:searchRetrieveResponse>
    <srw:version>1.1</srw:version>
    <srw:numberOfRecords>206</srw:numberOfRecords>
   - <srw:records>
     - <srw:record>
        <srw:recordSchema>oai</srw:recordSchema>
        <srw:recordPacking>xml</srw:recordPacking>
       - <srw:recordData>
         - <oai_dc:dc xsi:schemaLocation="http://www.imsglobal.org/xsd/imsmd_v1p2p4.xsd">
           - <oai:general>
             - <oai:title>
                 <oai:langstring xml:lang="nl">Fossiele slang zo lang als T-rex</oai:langstring>
               </oai:title>
             - <oai:catalogentry>
                 <oai:catalog>tag</oai:catalog>
               - <oai:entry>
                   <oai:langstring xml:lang="x-none">tag:www.kennisnet.nl,2009:Publication/3326</oai:langstring>
                 </oai:entry>
               </oai:catalogentry>
               <oai:language>nl</oai:language>
             - <oai:description>
               - <oai:langstring xml:lang="nl">
                   In Colombia is de grootste slang ook gevonden. Het fossiel is naar schatting dertien meter lang. De slang is als oerthermometer gebruikt voor de tropen van zo'n 60 miljoen jaar geleden. Het was toen nuim warmer dan tegenwoordig. Worden de kwetsbare tropen nog warmer als de aarde verder opwarmt?
                 </oai:langstring>
               </oai:description>
             - <oai:keyword>
                 <oai:langstring xml:lang="nl">Natuur</oai:langstring>
               </oai:keyword>
             - <oai:keyword>
                 <oai:langstring xml:lang="nl">aarde & klimaat</oai:langstring>
               </oai:keyword>
             - <oai:keyword>
```

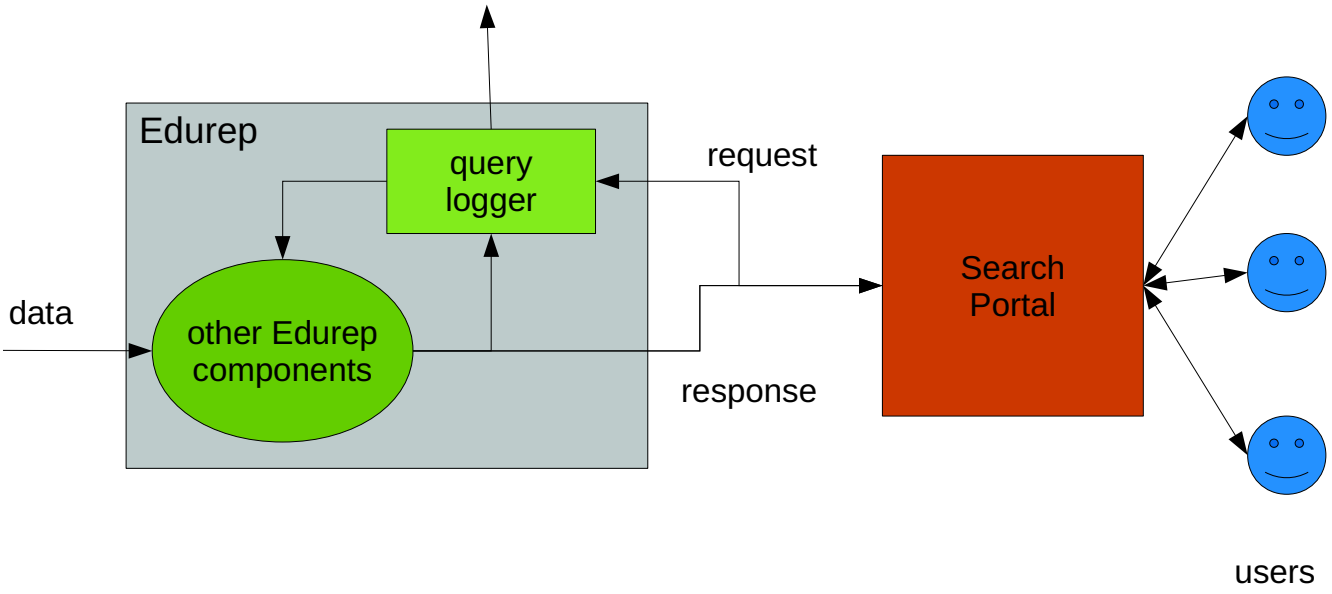

Search Portals





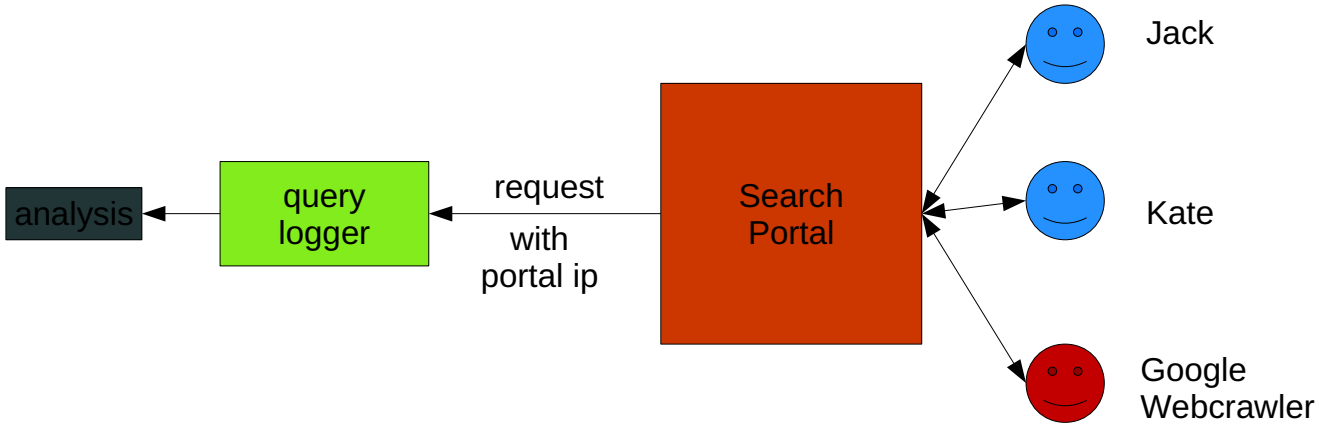
Query Logging

<datetime> <ip> <response size> <response time> <protocol> <query>



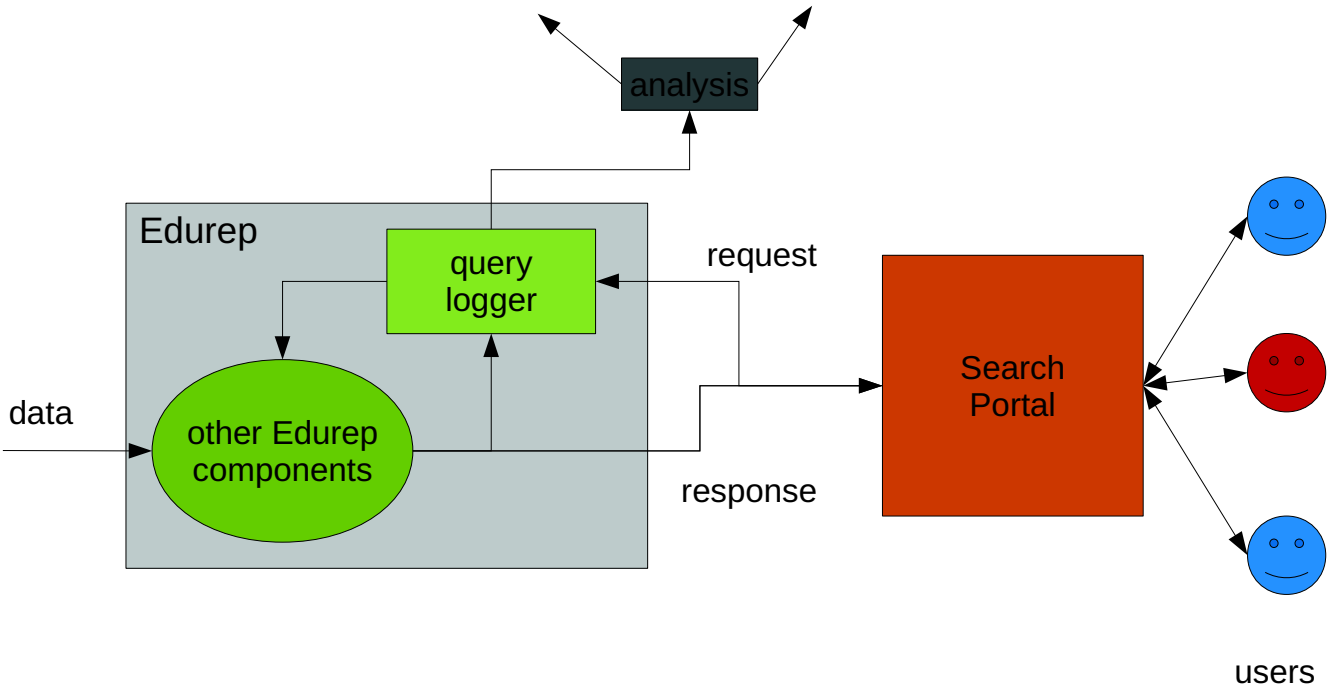
Problem

- Human and automated users on search portals.
  - extra traffic
- At Edurep level, regular means of user type distinction cannot be used.
  - no reliable user behaviour



Goal

- Distinguish automated and user generated queries

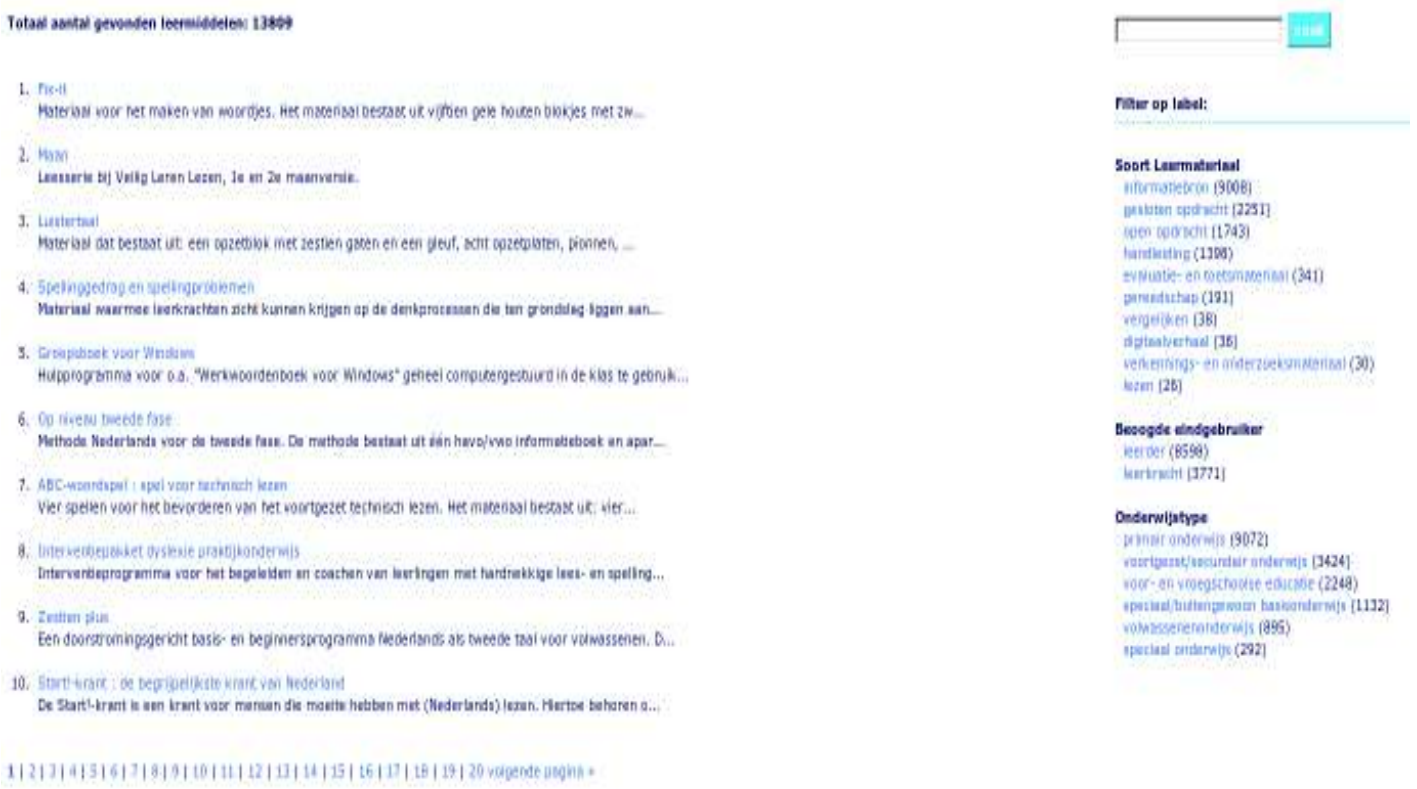


Hyperlinks

- Webcrawlers automatically follow hyperlinks.
- 4 types:
  - **search link**: retrieve a first page resultset
  - **pagination link**: retrieve another resultset page
  - **result link**: retrieve a specific record
  - **facet link**: retrieve the nr of records for that facet
- Each link type has a specific query pattern.

Search Link

retrieve a first page resultset



Search Link

retrieve a first page resultset

- startRecord == 1

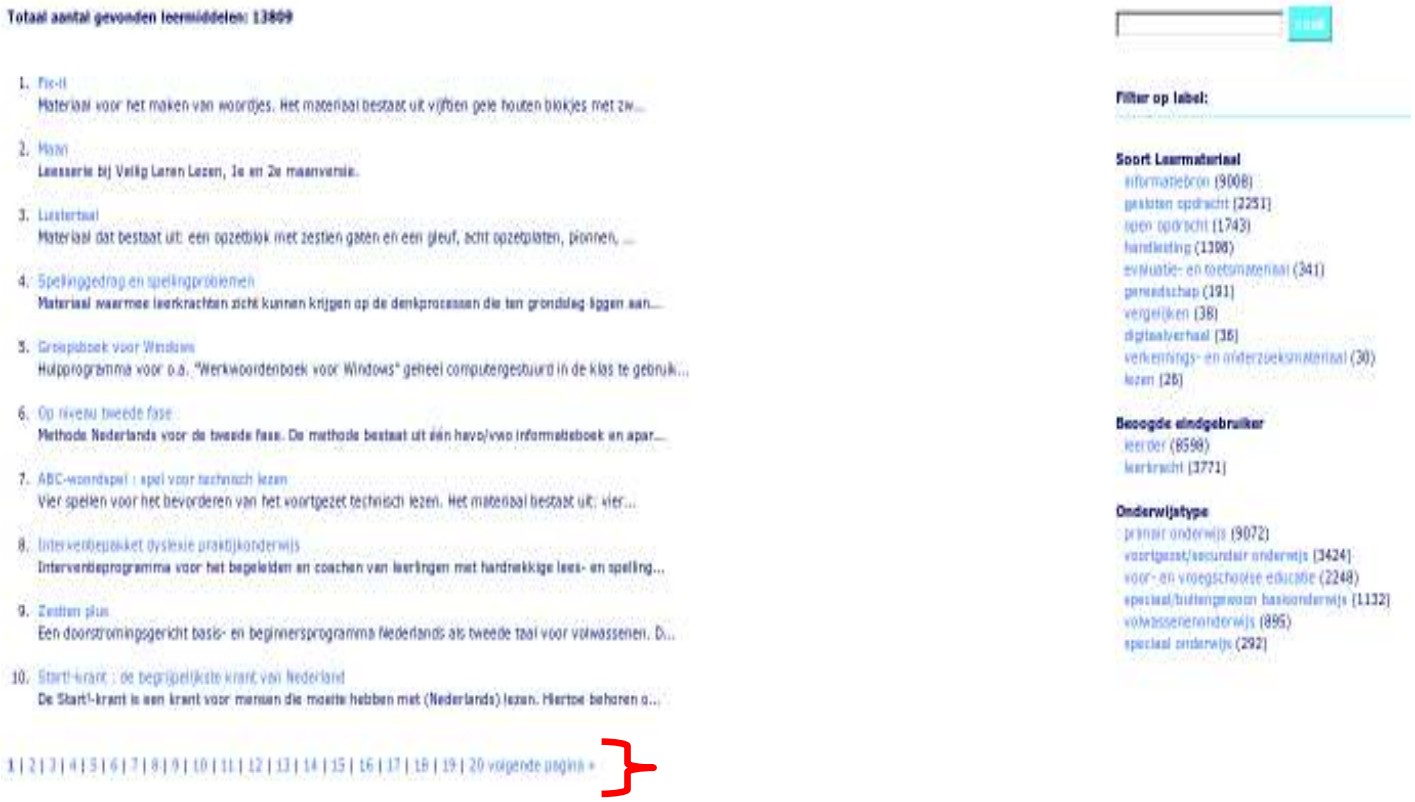
  OR
- startRecord == null (using default 1)

  AND
- maximumRecords >= 5

  OR
- maximumRecords == null (using default 10)

Pagination Links

retrieve another resultset page



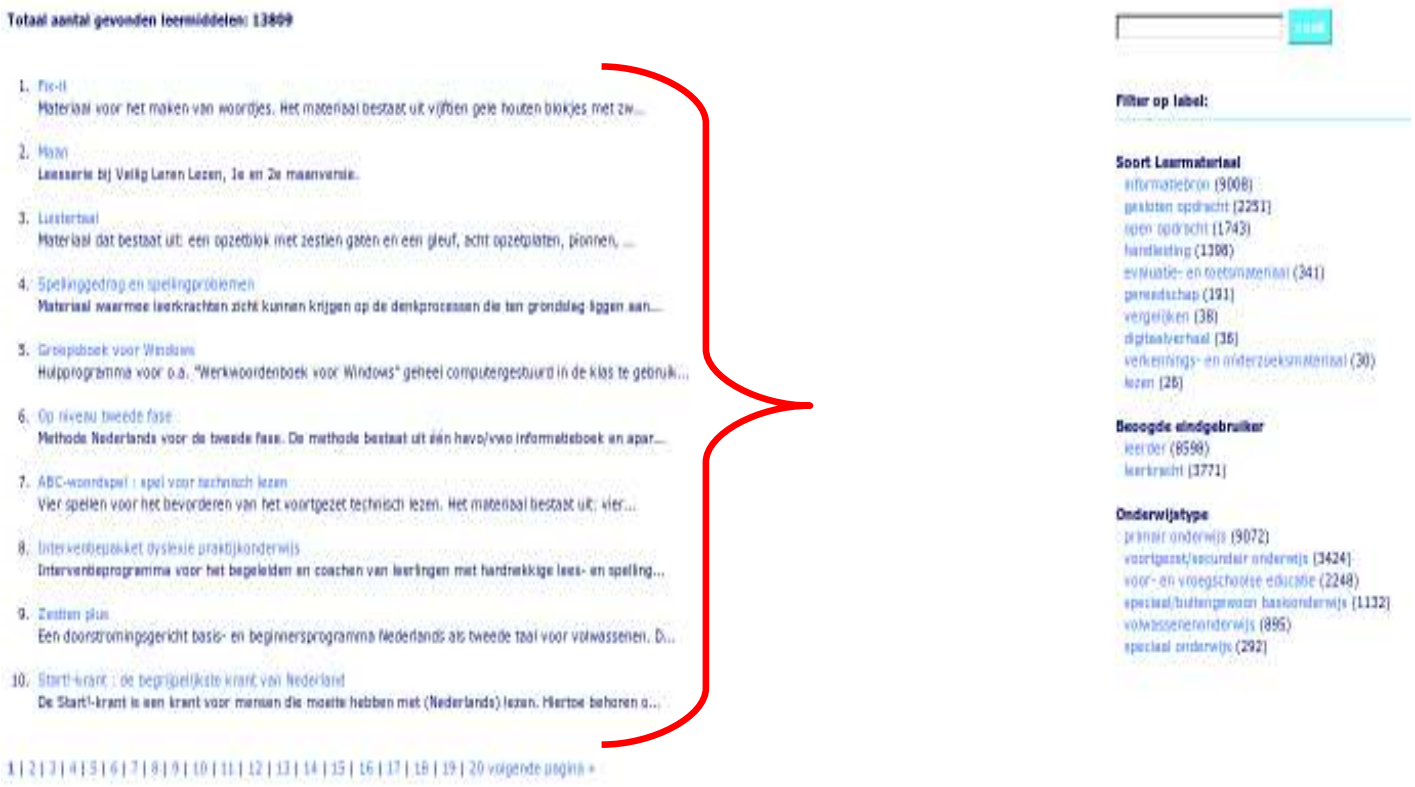
Pagination Links

retrieve another resultset page

- startRecord > 1

Result Links

retrieve a specific record



Result Links

retrieve a specific record

- startRecord == 1

  OR

- startRecord == null (using default 1)

  AND

- query on unique record aspect (eg. catalogentry)

Facet Links

retrieve the amount of records for that facet

Facet Links

retrieve the amount of records for that facet
(using facet index)

- query

  AND
- x-term-drilldown

```
<srw:extraResponseData>
  <dd:drilldown xsi:schemaLocation="http://meresco.org/namespace/drilldown http://m
    <dd:term-drilldown>
      <dd:navigator name="lom.educational.context.value">
        <dd:item count="769">VO</dd:item>
        <dd:item count="241">PO</dd:item>
        <dd:item count="181">BVE</dd:item>
        <dd:item count="30">VVE</dd:item>
        <dd:item count="26">primaire onderwijs</dd:item>
        <dd:item count="21">voortgezet onderwijs</dd:item>
        <dd:item count="12">beroepsonderwijs en volwaseneneducatie</dd:item>
        <dd:item count="9">speciaal onderwijs</dd:item>
        <dd:item count="9">speciaal basisonderwijs</dd:item>
        <dd:item count="7">other</dd:item>
      </dd:navigator>
      <dd:navigator name="lom.rights.cost">
        <dd:item count="951">no</dd:item>
        <dd:item count="58">yes</dd:item>
      </dd:navigator>
      <dd:navigator name="lom.general.aggregationLevel">
        <dd:item count="931">1</dd:item>
        <dd:item count="153">2</dd:item>
        <dd:item count="78">4</dd:item>
        <dd:item count="38">3</dd:item>
      </dd:navigator>
      <dd:navigator name="lom.technical.format">
        <dd:item count="741">text/html</dd:item>
        <dd:item count="584">image/jpeg</dd:item>
        <dd:item count="318">video/x-ms-asf</dd:item>
        <dd:item count="31">application/zip</dd:item>
        <dd:item count="24">application/pdf</dd:item>
        <dd:item count="16">non-digital</dd:item>
        <dd:item count="5">application/msword</dd:item>
        <dd:item count="4">text/xml</dd:item>
        <dd:item count="2">application/octet-stream</dd:item>
        <dd:item count="2">image/gif</dd:item>
      </dd:navigator>
    </dd:term-drilldown>
  </dd:drilldown>
</srw:extraResponseData>
```

Facet Links

retrieve the amount of records for that facet
(not using facet index)

- startRecord == 1 OR null (using default 1)

  AND
- maximumRecords == 0 OR 1*

  AND
- ! query on unique record aspect (eg. catalogentry)

  foreach
- query=lom.field=value

  (to retrieve the nr of records for that value)

Dataset

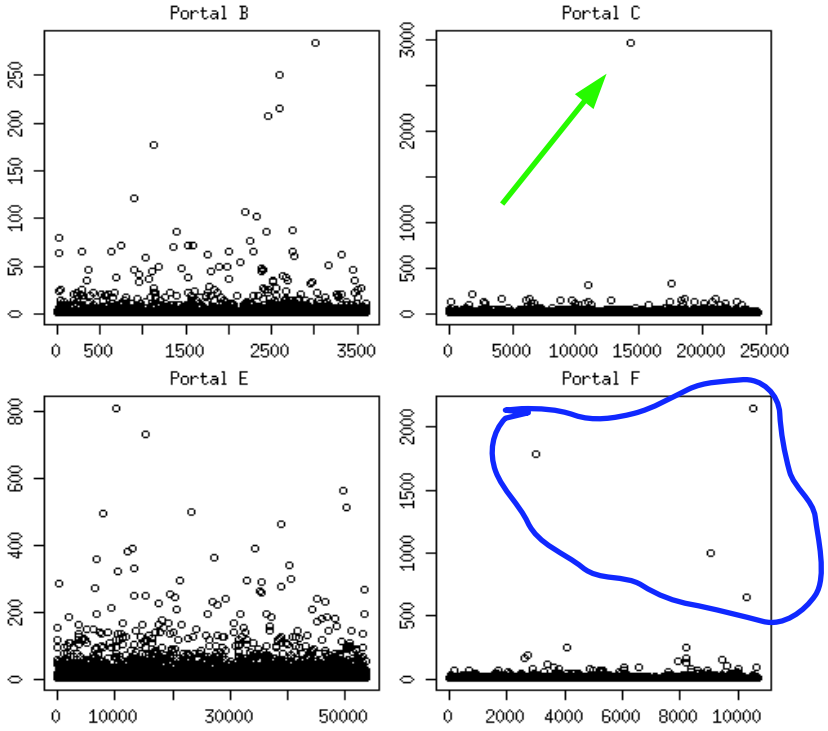
- Januari 2010 logs
- 6 largest portals

portal	requests
A	41.690
B	126.340
C	1.293.902
D	48.841
E	232.341
F	82.527
other	63.266
total	1.888.907

- 5 variables from each query
  - portal ip
  - startRecord
  - maximumrecords
  - search query hash
  - unique request boolean

Search Link

retrieve a first page resultset

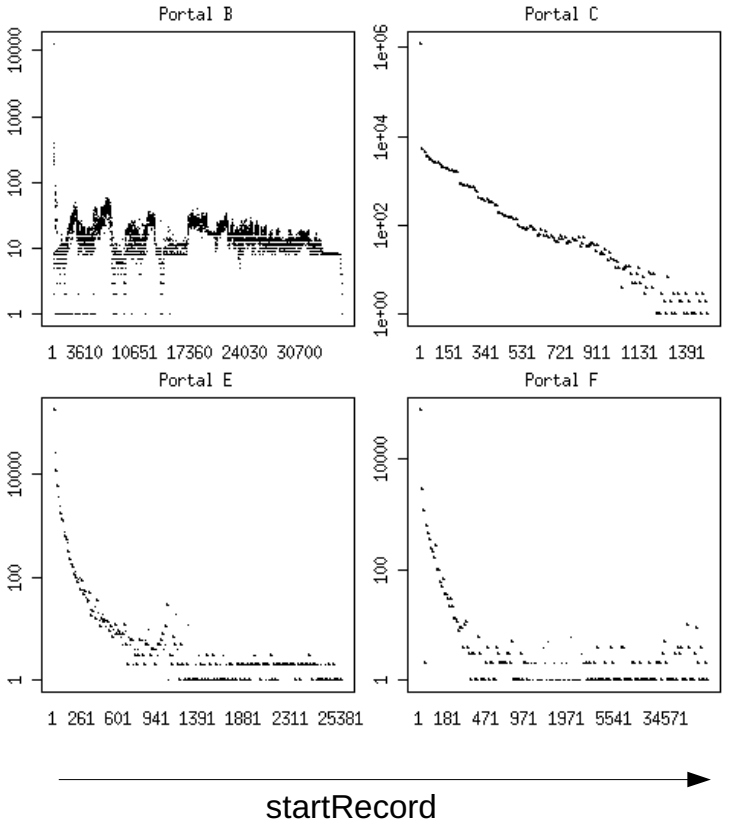


- counting unique query occurence
- start page, how much automatic?
- preset portal queries

  lom.educational.context=PO

Pagination Links

retrieve another resultset page



- mapping unique startRecord occurence

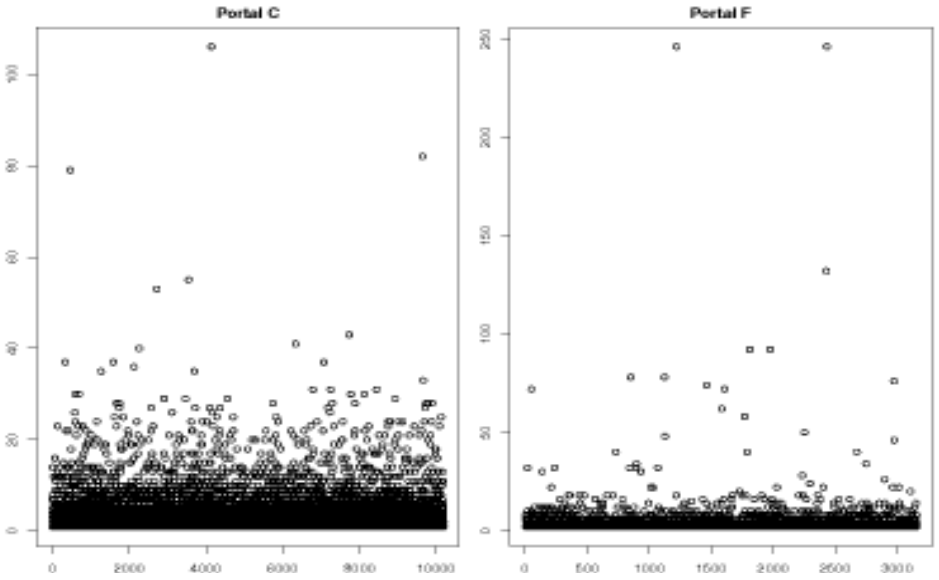
automatic querying behaviour

normal querying behaviour

Result Links

retrieve a specific record

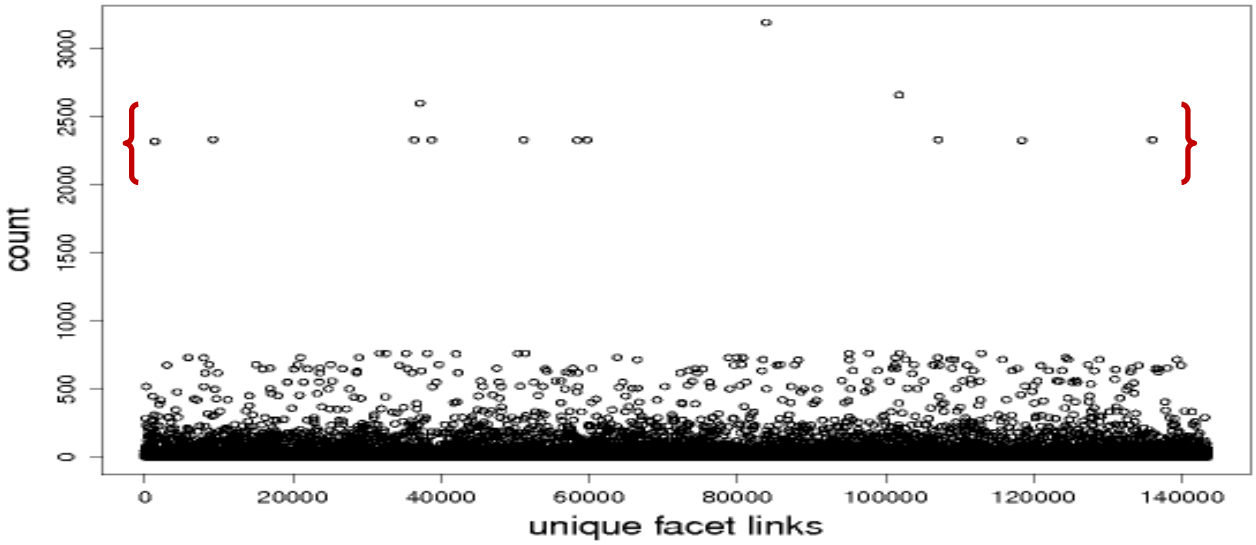
- not much result querying going on
- Edurep results are dynamic



Facet Links

retrieve the amount of records for that facet
(not using facet index)

- only portal c used this link type
- points corresponded with facet links on page



portal C

Observation Summary

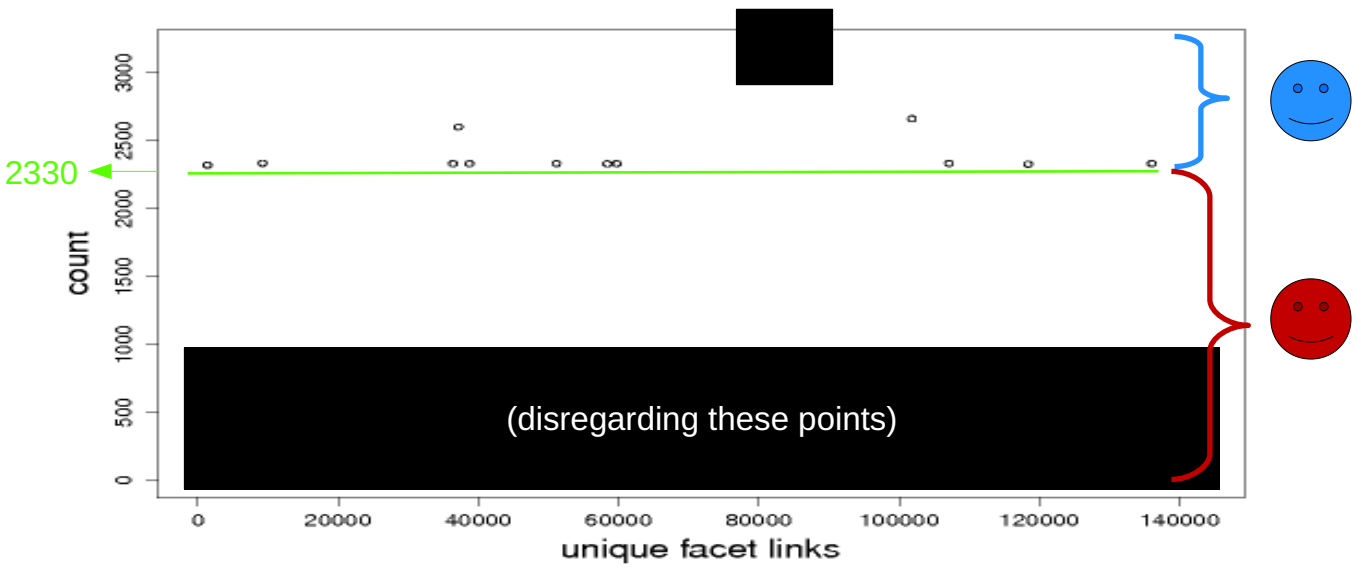
Automated querying indicated in:

- Facet links
- Pagination link

The question is how much...

Portal C's Facet Links

- human generated, but inefficient, distorting results
- 2330 x 12



Pagination Filtering

with maximumsRecord = 10
startRecord > 200 = resultpage 20

- Crude method:

  startRecord > 200 = automated (PAG1)

- Elegant method:

  filter pagination ranges (PAG2)

Pagination Ranges

first – 1 – 2 – 3 – 4 – 5 – last

1. query=dog&startRecord=1&maximumRecords=10
2. query=dog&startRecord=11&maximumRecords=10
3. query=dog&startRecord=21&maximumRecords=10
4. query=dog&startRecord=31&maximumRecords=10
5. query=dog&startRecord=41&maximumRecords=10

A Pagination Range

- is a set of equal queries, except for startRecord
  - minimum of 10 queries
  - maximum startRecord value > 200
  - a startRecord difference of maximumRecords between each query

Pagination Results

portal	requests	crude (PAG1)	elegant (PAG2)
A	41.690	-15.237	-13.355
B	126.340	-105.026	-89.710
C	1.293.902	-15.255	-15.654
D	48.841	-47	-62
E	232.341	-1778	-1.815
F	82.527	-1.778	-205
total	1.825.641	-137.749	-120.801

- PAG2 did not take into account:
  - heads and tails outside the logs
  - first/previous/next/last page links

Results in Context

- PAG2's queries requested 13,3 Gb
- 6.6% of total requests

Immediate Findings

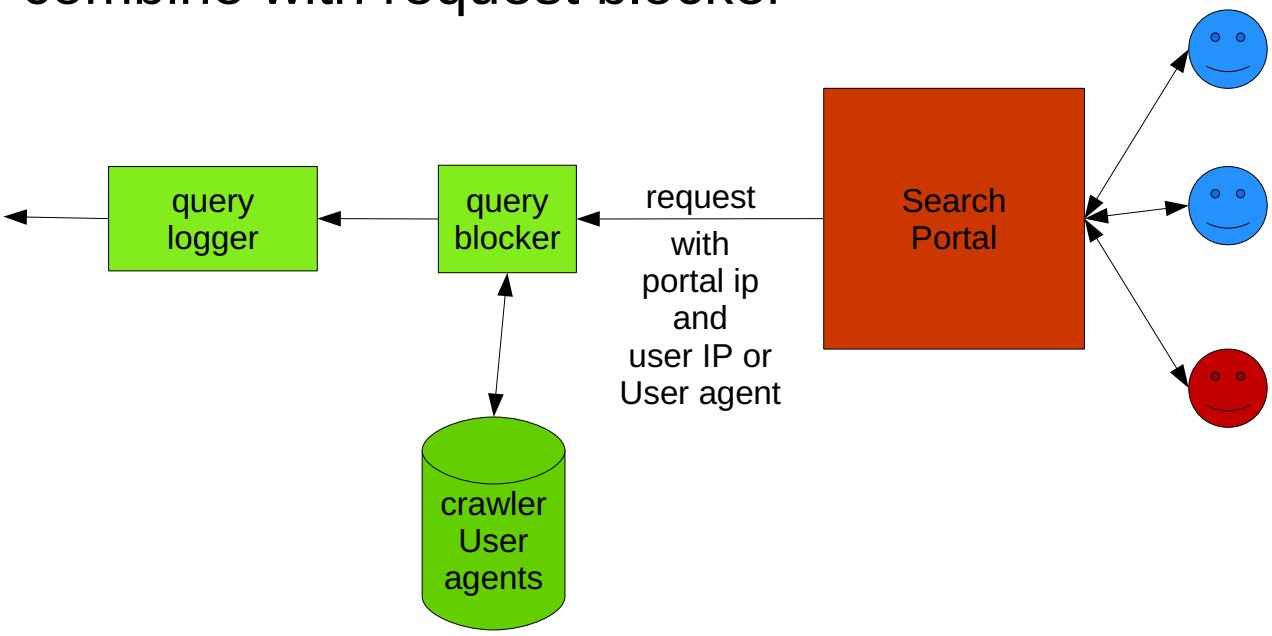
- Advise
  - blocking crawlers → Robot Exclusion Protocol
  - more efficient querying
- Administration
  - expanding & automating scripts
  - integrate in tooling

Research Improvement

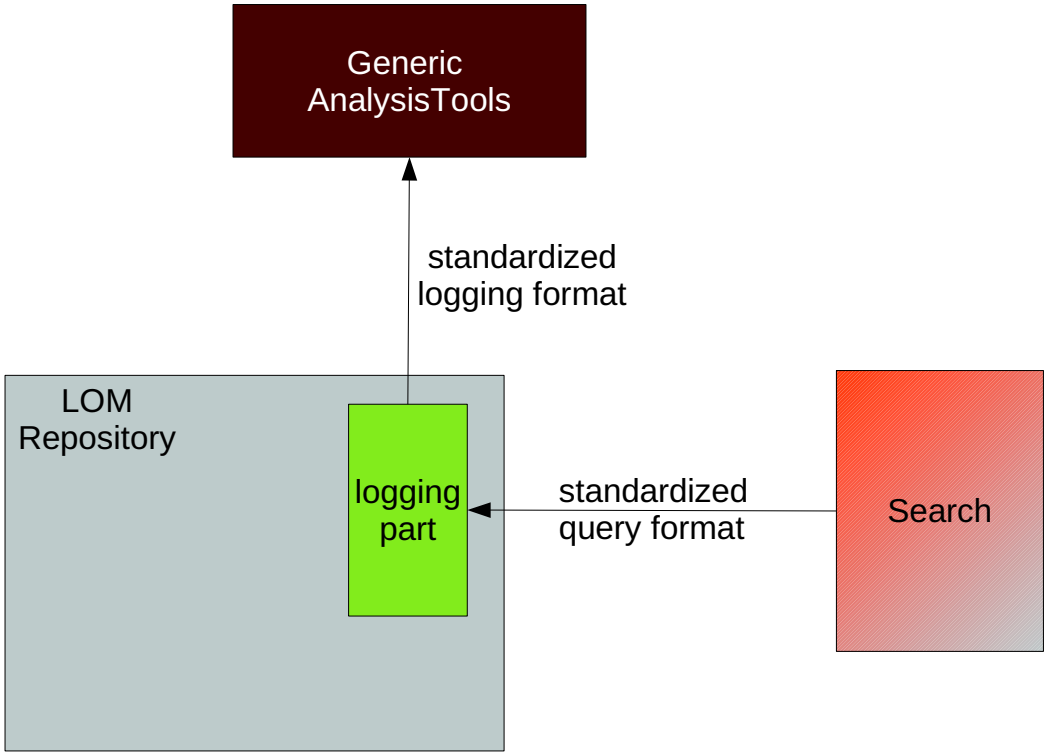
- use and compare with more logging
- detailed information on query
  - query={beer}AND{lom.general.title=abuse}AND {lom.technical.format=text/html}
- Use specific portal parameters in analysis
  - size and format of pagination
  - other search interaction links

Possible System Changes

- require user IP adress or User agent in query
- combine with request blocker



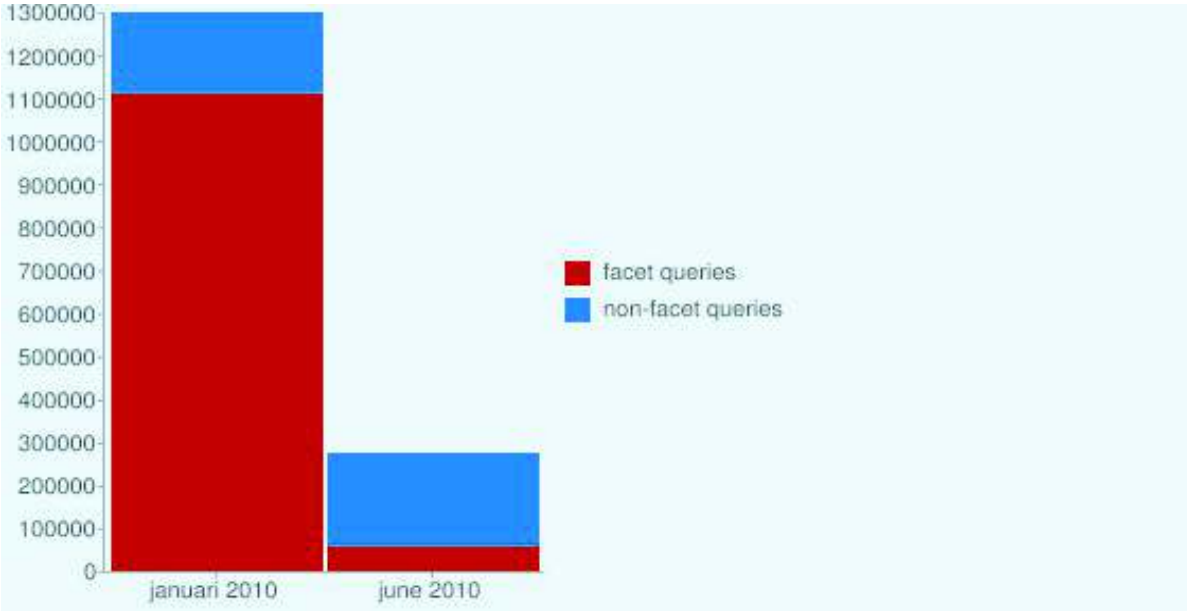
Generic Logging Tools



Some continued work...

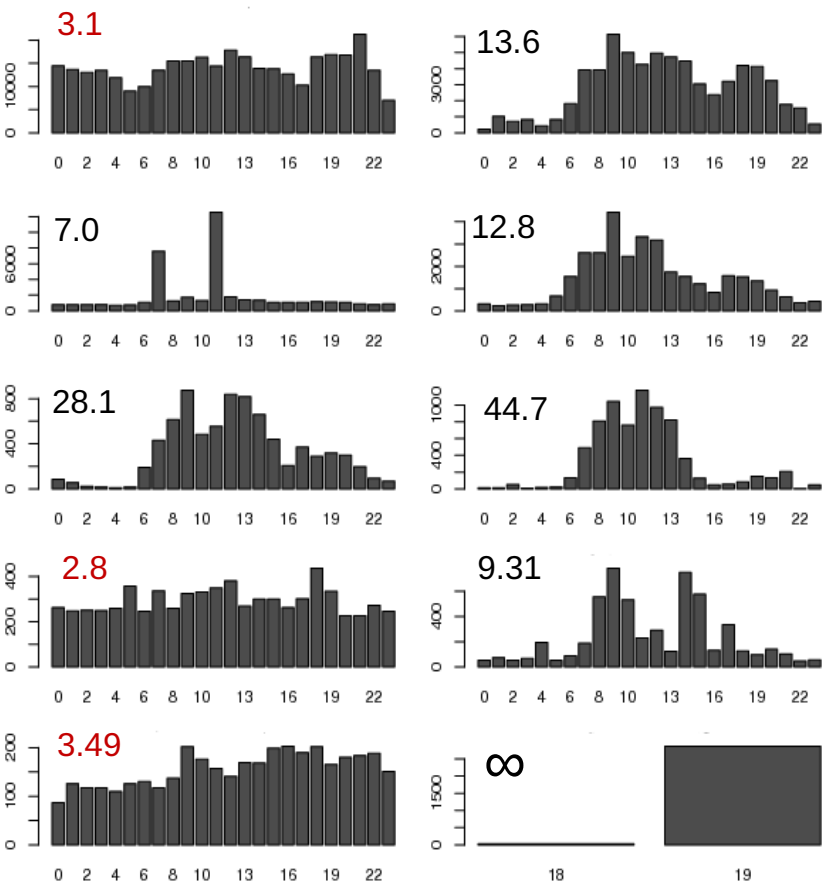
Advise Effect

- Results for consulting portal C on effective querying.

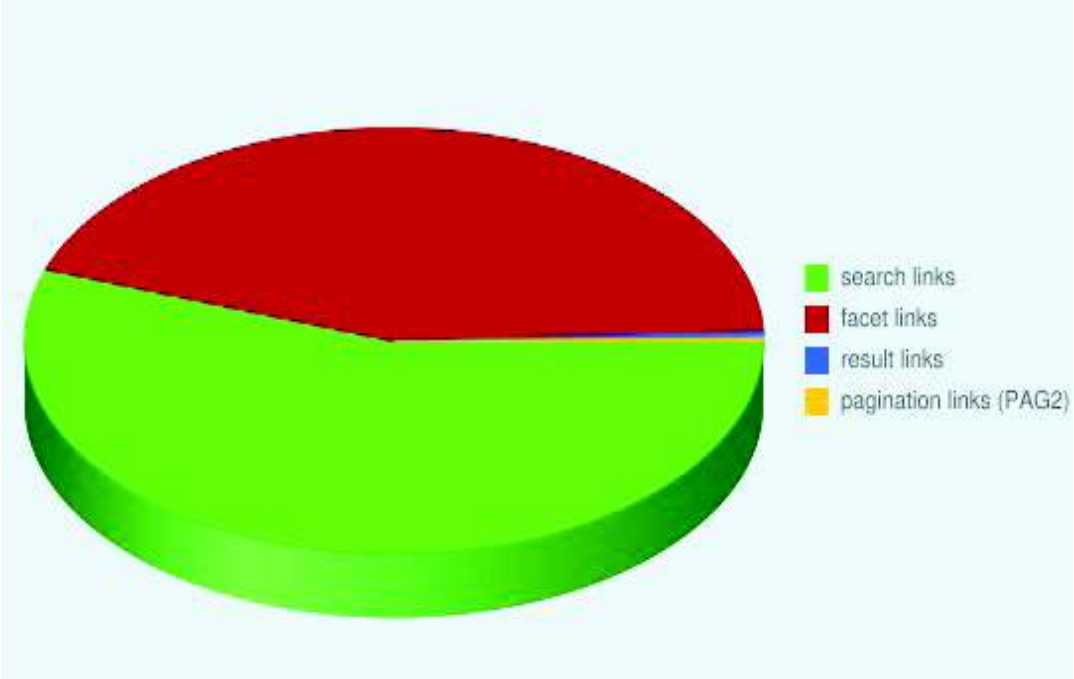


Querying, day and night

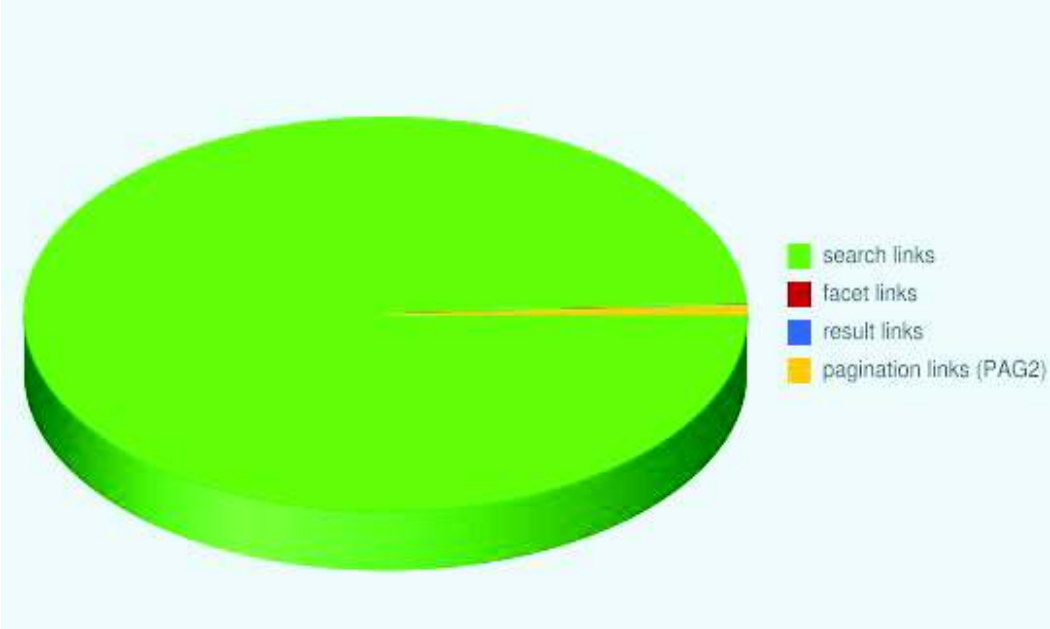
- Dutch content, so access from 1 timezone
- day when hour > 5:00 & < 0:00
- if ratio day / night < 5 automatic querying?



Combining Link Types
Portal F



Combining Link Types
Portal E



Combining Link Types
Portal C