Patent-Crawler

A real-time recursive focused web crawler to gather information on patent usage

HPC-AI Advisory Council, Lugano, April 2018

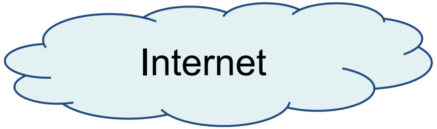
E. Orliac[1,2], G. Fourestey[2], D. Portabella[2], G. de Rassenfosse[2] (1: UniL, 2: EPFL)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

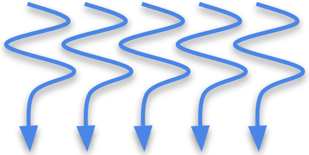
INFORMATIQUE SCIENTIFIQUE & SUPPORT APPLICATIF
SCITAS
SCIENTIFIC IT AND APPLICATION SUPPORT

UNIL | Université de Lausanne

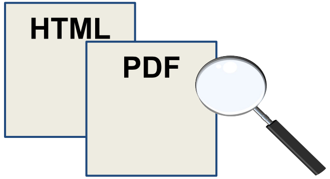
Crawler overview



Seed list: Set of URL to start from

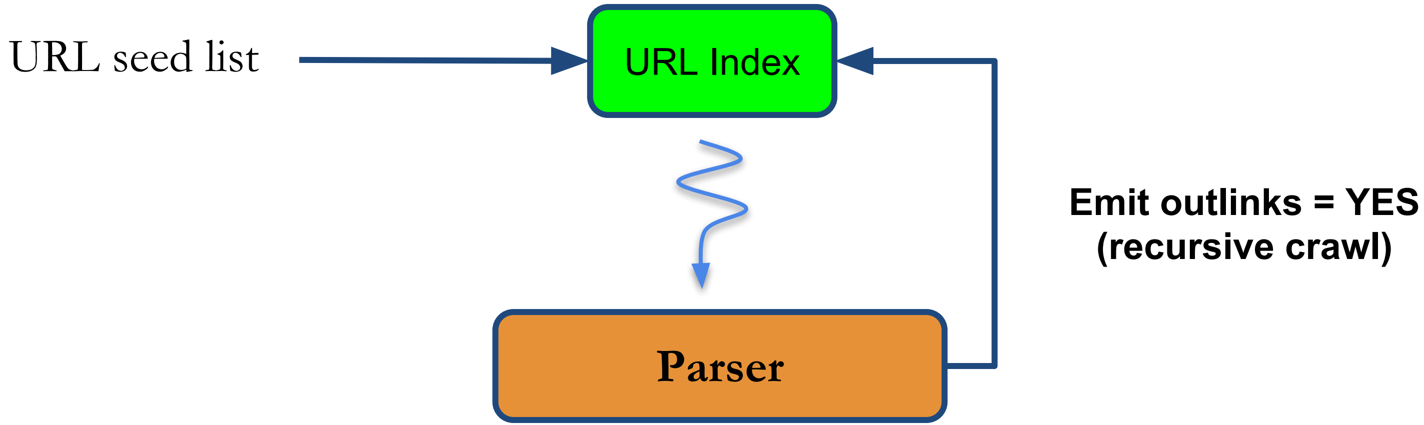


Fetcher: Downloads the documents



Parser: Scans the documents for targeted information (**focused** crawler)

Archiver: Create WARC files (typ.) from selected web pages



Several ways to act on the search space (URL frontier):

- Seed list
- URL filters
- Crawl depth
- Types of documents to consider
- Emit outlinks (if YES, recursive crawl)

Motivations

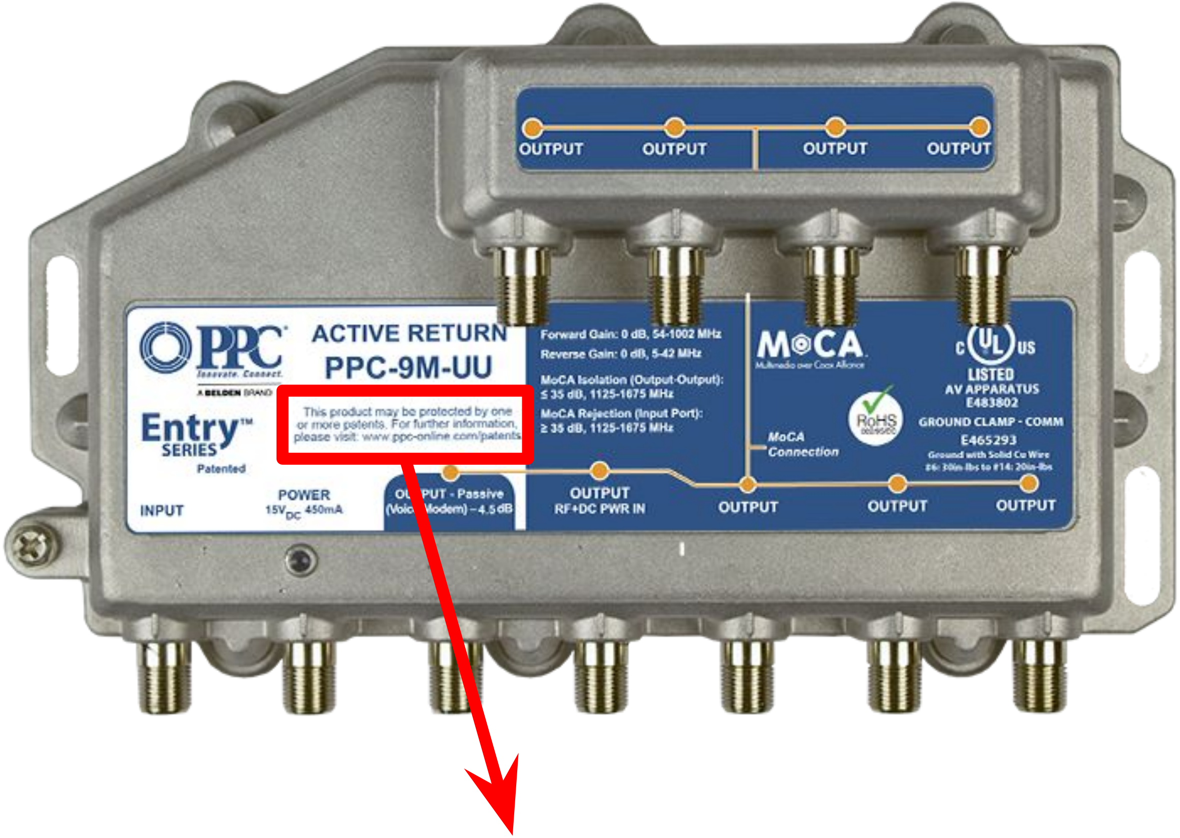
> **35 U.S. Code § 287 - Limitation on damages and other remedies; marking and notice**
>
> Patentees […] may give notice to the public that the [product] is patented, either by fixing thereon the word “patent” or the abbreviation “pat.”, together with the number of the patent, or by fixing thereon the word “patent” or the abbreviation “pat.” together with an address of a posting on the Internet, accessible to the public without charge for accessing the address, that associates the patented article with the number of the patent […].

Virtual Patent Marking (VPM) hence become a source of information wrt patenting with the major advantage that you do not need to physically access the product to see which patents are used.

Relevance of *patent* ↔ *product* information

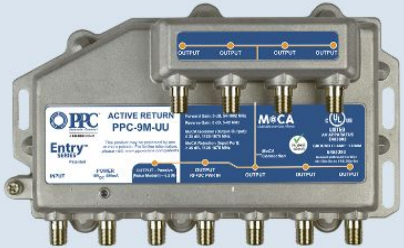
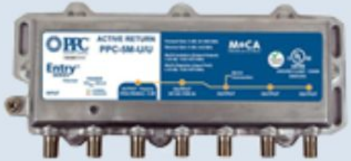
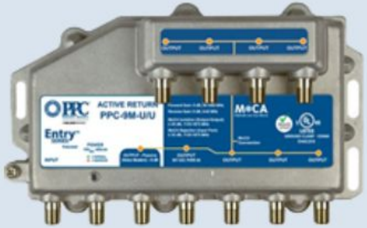
- Provides a direct link between innovation and market
- Hence a way to assess the role of science and technology on the economy
- The establishment of a *Patents* ↔ *Products* database, one of the holy grails of innovation research
- The Chair of Innovation and IP Policy (IIPP) of EPFL asked the Scientific IT and Application Support group (SCITAS, EPFL) to setup a focused web crawler

🌐 http://www.iproduct.io


OUTPUT
OUTPUT
OUTPUT
OUTPUT
PPC
Innovate. Connect.
A BELDEN BRAND
ACTIVE RETURN
PPC-9M-UU
This product may be protected by one or more patents. For further information, please visit: www.ppc-online.com/patents
Entry™
SERIES
Patented
Forward Gain: 0 dB, 54-1002 MHz
Reverse Gain: 0 dB, 5-42 MHz
MoCA Isolation (Output-Output):
≤ 35 dB, 1125-1675 MHz
MoCA Rejection (Input Port):
≥ 35 dB, 1125-1675 MHz
MoCA
Multimedia over Coax Alliance
c UL us
LISTED
AV APPARATUS
E483802
RoHS
MoCA
Connection
GROUND CLAMP - COMM
E465293
Ground with Solid Cu Wire
#6: 30in-lbs to #14: 20in-lbs
INPUT
POWER
15VDC 450mA
OUTPUT - Passive
(Voice Modem) – 4.5 dB
OUTPUT
RF+DC PWR IN
OUTPUT
OUTPUT
OUTPUT

http://www.ppc-online.com/patents http://www.ppc-online.com/patents

	Virtual patent marking	Products	Patent numbers

ENTRY SERIES® Gen II	
	
	
		"PPC-5M-xxxx"
		OR "PPC-9M-xxxx"
		(e.g., PPC-5M-UU;
		PPC-5M-UUPI;
		PPC-5M-UUPS;
		PPC-9M-UU;
		PPC-9M-UUPI;
		PPC-9M-UUPS)
			7,544,086
			8,286,209
			8,356,322
			9,516,376
			D751,509
			D756,935

How to access VPM information

- Crawl the web
- Use publicly available general crawl datasets (such as CommonCrawl)

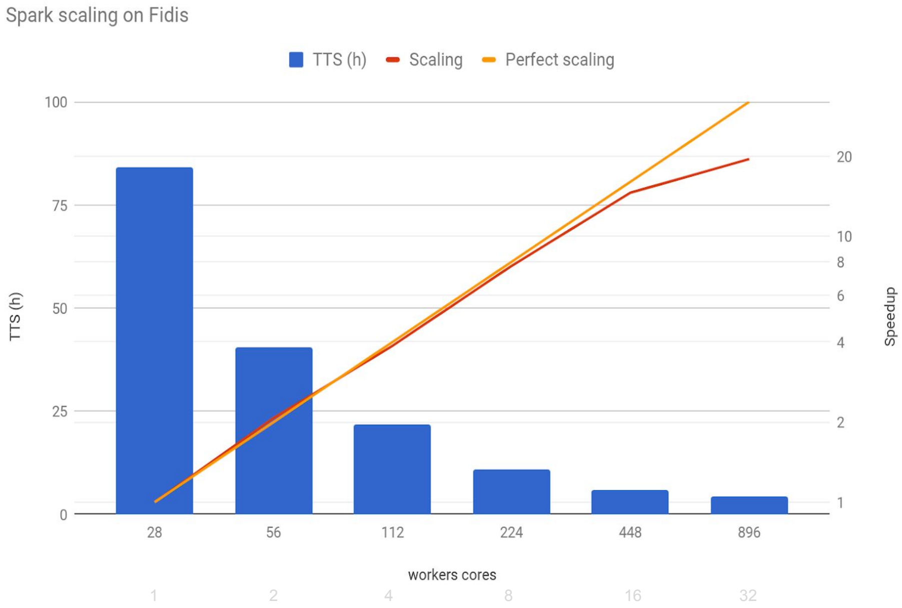
Pros and cons of both approaches

Crawling the web	Public crawl data
Pros	
+ Fully independent	+ No need to run a crawler
+ Fully customizable	+ Reusable (non-focused)
+ Reduced datasets (focused crawler)	
Cons	
- Need to build, tune and operate a crawler	- Need to download huge datasets or process on data host
- Need an adequate infrastructure for large/massive crawls	- Need an adequate infrastructure to process TB of data
- Need to handle politeness	- Fully dependent on data provider
- Risk of being blacklisted	

The *vpmfilter* tool from IIPP

- Purpose: scan CommonCrawl (CC) data in search of VPM information
- Based on Spark

- Scales nicely
- Less than 1 day processing on 4 nodes on fidis@EPFL to process a monthly crawl of ~30 TB
- About 4.3 hours on 896 cores

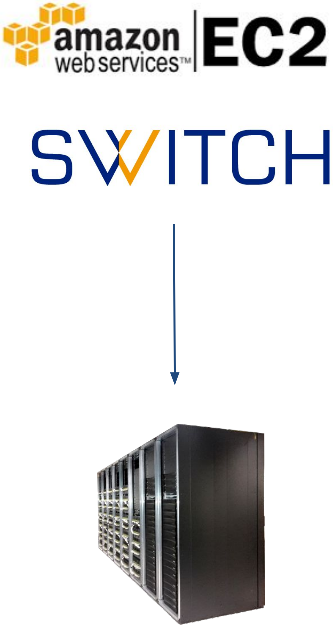


G. Fourestey et al., Big Data on HPC Clusters Tracing innovations in the marketplace, Fidis Unveiling, June 15th 2017

Dataflow

But…

- Took 3 weeks to download 30 TB of data from Amazon @30 MB/s



Hence the motivations to set up a focused crawler

- Massively reduce the amount of data to transfer (~98%)
- However the actual full regex (keyword + number) in the Parser bolts has a major impact on the crawler performances

Hybrid approach

- Crawler on patent keyword only (fast) on cloud
- *vpmfilter* on HPC on patent number on HPC clusters

G. Fourestey et al., Big Data on HPC Clusters Tracing innovations in the marketplace, Fidis Unveiling, June 15th 2017

Focused crawler design

Setup:

- Apache Storm http://storm.apache.org/
- StormCrawler http://stormcrawler.net/
- Elasticsearch https://www.elastic.co/

Goal:

- Identify VPM pages
- Archives VPM pages into WARC files for further analysis

Implementation:

- Dual regex: **patent-keyword + patent-number** (in the Parsers)

Does exactly the same parsing as the *vpmfilter* tool

WARC-Target-URI: https://www.dpreview.com/articles/6776074667/panasonic-announces-lumix-dmc-fz200-superzoom-with-constant-f2-8-lens

```
<div class="widget minorArticlesWidget"><div class="widgetTitle">Latest articles</div><div class="widgetContent"><div class="article"><div class="image"><img
src="https://4.img-dpreview.com/files/p/E~C76x0S437x437T72x72~articles/7551928919/readers-choice-midrange-ilc-2017.jpeg"></div><div class="title"><a
href="https://www.dpreview.com/articles/1838050609/have-your-say-best-mid-range-ilc-of-2017" target="_self">Have your say: Best mid-range ILC of 2017</a></div><div class="summary"><p>This year saw
several cameras released in the mid-range ILC class, from full-frame DSLRs to super-compact APS-C mirrorless models. Take a look for a reminder of the key mid-range ILCs released in 2017, and don't forget to vote
for your favorites.</p></div><div class="info"><span class="time">Dec 16, 2017</span><a class="comments"
href="https://www.dpreview.com/articles/1838050609/have-your-say-best-mid-range-ilc-of-2017#comments">26</a></div></div><div class="article"><div class="image"><img
src="https://2.img-dpreview.com/files/p/E~C76x0S437x437T72x72~articles/5933511531/readers-choice-entry-level-ilc-2017.jpeg"></div><div class="title"><a
href="https://www.dpreview.com/articles/1800375418/have-your-say-best-entry-level-ilc-of-2017" target="_self">Have your say: Best entry-level ILC of 2017</a></div><div class="summary"><p>The most
important camera you'll ever own is the first one you buy. This year was relatively quiet on the entry-level ILC front, but the quality of the cameras released in this market segment was universally excellent.
[…]
=%7b%22st%22%3a%22dpreview%22%7d&quot;,&quot;loadAfter&quot;:&quot;windowOnLoad&quot;,&quot;daJsUrl&quot;:&quot;https://images-na.ssl-images-amazon.com/images/G/01/adFeedback/Feedbac
k-NA/feedback-us._CB315238478_.js&quot;}" style="height: 250px;"></div><div class="article"><div class="image"><img
src="https://3.img-dpreview.com/files/p/E~C116x0S685x685T72x72~articles/3295932675/359205269_c16aa03cf9_o.jpeg"></div><div class="title"><a
href="https://www.dpreview.com/news/3295932675/canon-patents-400mm-f5-6-catadioptric-mirror-lens" target="_self">Canon patents 400mm F5.6 catadioptric &#39;mirror&#39; lens</a></div><div
class="summary"><p>Canon might be planning to bring catadioptric 'mirror' lenses back from the dead. A new Canon patent spotted in Japan describes a 400mm F5.6 catadioptric lens that would use a variable
density &lsquo;electrochromic&rsquo; filter to 'stop down' the lens.</p></div><div class="info"><span class="time">Dec 15, 2017</span><a class="comments"
href="https://www.dpreview.com/news/3295932675/canon-patents-400mm-f5-6-catadioptric-mirror-lens#comments">171</a></div></div><div class="article"><div class="image"><img
src="https://1.img-dpreview.com/files/p/E~C122x0S1000x1000T72x72~articles/7982593338/M10handsonDxO.jpeg"></div><div class="title"><a
href="https://www.dpreview.com/news/7982593338/dxomark-the-full-frame-leica-m10-is-on-par-with-the-best-aps-c-sensors" target="_self">DxOMark: The full-frame Leica M10 is &#39;on par&#39; with the
best APS-C sensors</a></div><div class="summary"><p>DxOMark just finished their review of the Leica M10 sensor, and while it outperforms almost every other dig
[…]
```

Post-processing required to extract useful information, if any.

Initial tests at CSCS

CSCS kindly reserved a node for testing the crawler, but:

- Too much pressure on DNS server
- Security/ethical concerns with the nature of visited sites

Moved to SWITCH

- Small server for prototyping/investigating the crawler (4 CPUs, 16 GB RAM, 100 GB SSD)
- A priori no concern with the nature of crawled sites
- High performance DNS servers
- Happy to provide support in case of troubles

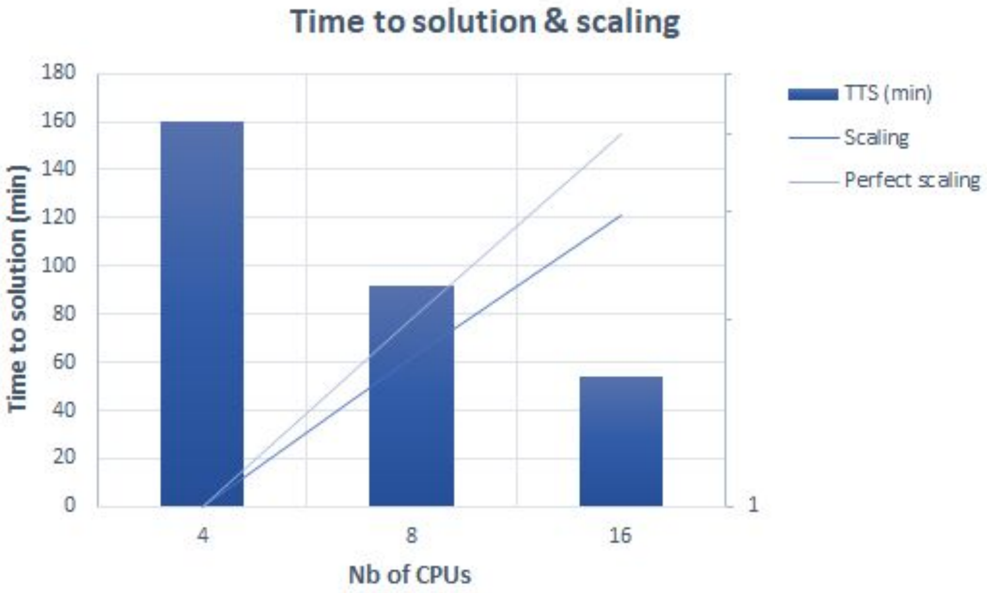
Tuning the crawler

Many factors will impact the crawling performances (other than computing resources available):

- The distribution of the URLs
- Queuing design of the Fetcher bolt
- Number of fetching threads
- Number of threads per queue
- Politeness of the crawler (robot.txt)
- Parallelism in the topology
- Parser efficiency
- Indexing system (in particular for recursive crawls)
- …
- + all what is not under control and subject to changes...

Performance scaling on AWS

- Selected ~2 M URLs from 56 segments of CC data
- Only kept domains with ≤ 25 URLs
- Non-recursive crawl on 3 AWS EC2 “compute” instances



Test case 1: *vpmfilter* vs crawler (keyword only)

Dataset:

- 28 segments of the CC dataset Nov. 2017, ~28 GB of zipped WARC files
  (28 as per the number of cores per node on fidis@SCITAS)
- 1,130,515 URLs
- Run the crawler in non-recursive mode (no new URL discovery) to control the search space

Goals:

- Estimate crawler efficiency and cost against the *vpmfilter*
- **Comparison is not perfect: between the CC crawl and ours, several weeks have passed, crawler setups are different, post-processing, ...**

Test case 1: *vpmfilter* vs crawler (keyword only)

Number of positive matches:

- *vpmfilter*: 12,569 (ref.); crawler: 12,789

Common matches: 9,970
Missed matches: 2,599

ES status distribution of URLs with mismatches:

"Missing" distribution	Number
REDIRECTION (not followed)	1,356
FETCHED	919
ERROR	305
Not in ES index (filtered somehow)	19

Indicative times to solution

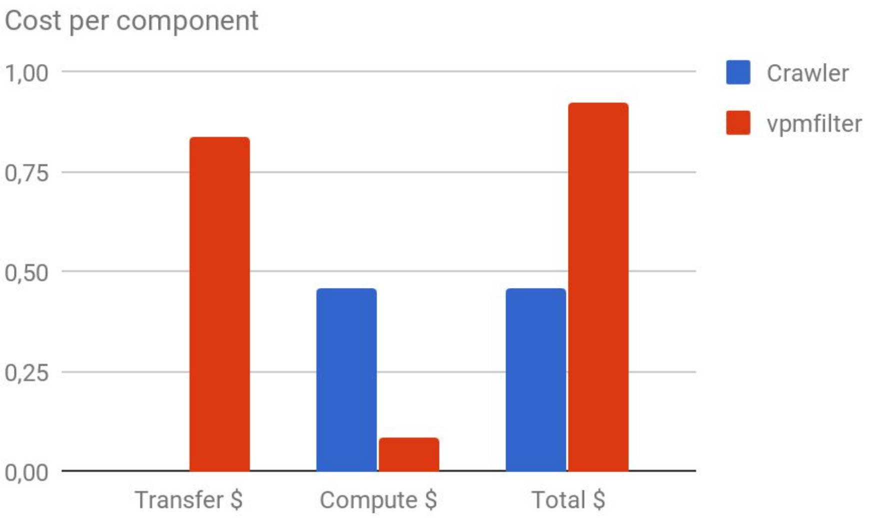
The crawler was run on SWITCH, *vpmfilter* on 1 node of fidis@EPFL

Crawler			**vpmfilter**	
PatentESSeedInjector	~3 min		Download CC data	~ 53 min
PatentCrawlTopology	~97 min		vpmfilter run (scaled to 4 cores)	~ 70 min
Total	**~100 min**		**Total**	**~123 min**

TCO comparison

The crawler was run on SWITCH, *vpmfilter* on 1 node of fidis@EPFL

	Crawler	vpmfilter
Host	SWTICH	Fidis@EPFL
Transfer (/GB) $	0,000	0,030
Core hour $	0,069	0,018
TTS 1 core (hours)	6,667	4,667



Roadmap towards deployment

- Access fundings for both manpower and compute resources
- Move to a multi-node setup
- EPFL MSc project proposal was published to deploy a large-scale crawler https://goo.gl/lB6Ibd
- Collaboration with StormCrawler to setup/develop a more efficient indexing/queuing system for large recursive crawls

Conclusions and outlook

- A prototype is ready for crawling VPM information
- Modularity: the system can easily adapted to other needs
- Scalability: the architecture used is similar to the one a distributed system would require
- The targeted amount of data to be crawled shall be carefully estimated before turning to production mode
- Also important to consider “hybrid” solutions

Acknowledgements

- CSCS (Colin J. McMurtrie & Hussein N. Harake) for providing exclusive access to one of their test node and providing support
- Julien Nioche for providing StormCrawler and support
- David Portabella from IIPP for continuous discussion about the crawler and for providing the *vpmfilter*
- Simon Leinen and Jens-Christian Fischer at SWITCH for exchanging on our crawler project

Thank you for your attention!