European Language Resource Coordination
Connecting Europe Facility

TILDE

European Commission

SMART 2019/1083

SMART 2019/1083 – TASK 6 ASSESSMENTS, TRIALS AND EVALUATIONS

MULTILINGUALISM SCORING TOOL

RINALDS VĪKSNA, 25.11.2021.

Connecting Europe Facility

MOTIVATION

Making European websites more multilingual is one of the targets of the Connecting Europe Facility Automated Translation (CEF AT). In order to monitor this goal, alongside other possible solutions, CEF AT needs a methodology and a tool to assess the degree of multilingualism of a web site is needed. The tool should be fully or semi-automatic and easy to use. This document covers the proposed acceptance criteria for and assessment under SMART 2019 1083 Task 6.

The emphasis is on the openness of the tool, as public administration and small business need to run it by themselves to evaluate the degree in which their website is multilingual. The tool can have proprietary components, but the main functionality of the tool should not depend on them.

Motivation

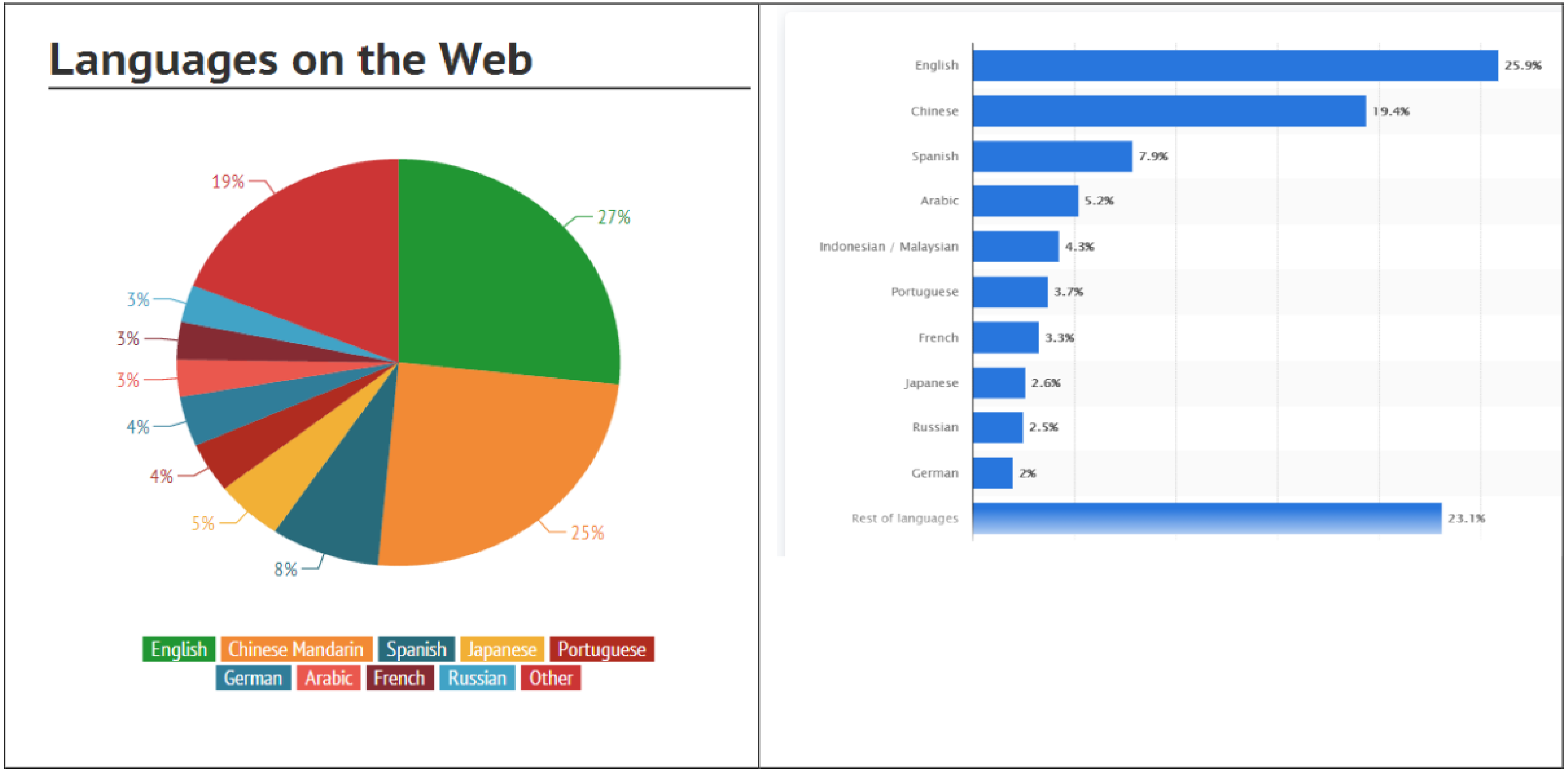
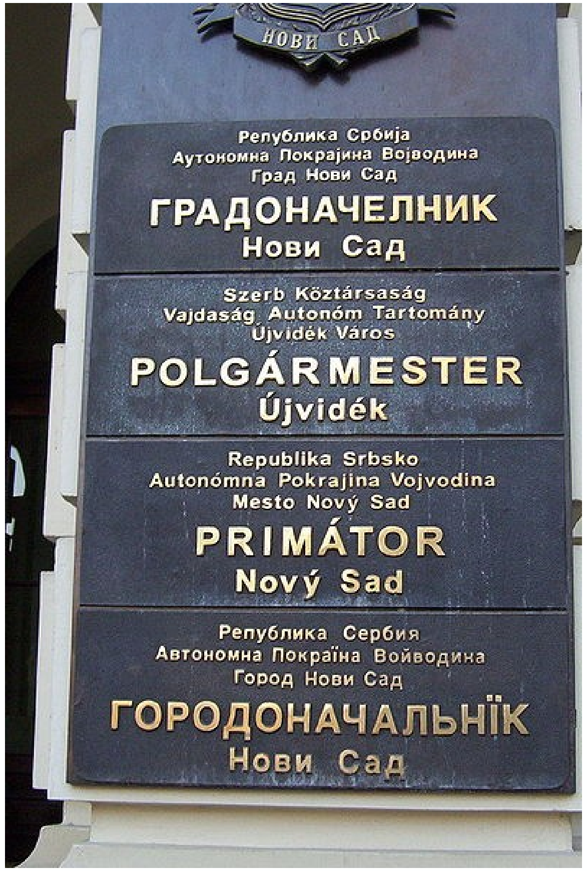


Figure 1. On left - Language of the Web (Web Language Diversity in Numbers | Language & Technology | Globalme); on right - Most common languages used on the internet 2020 (• Internet: most common languages online 2020 | Statista)

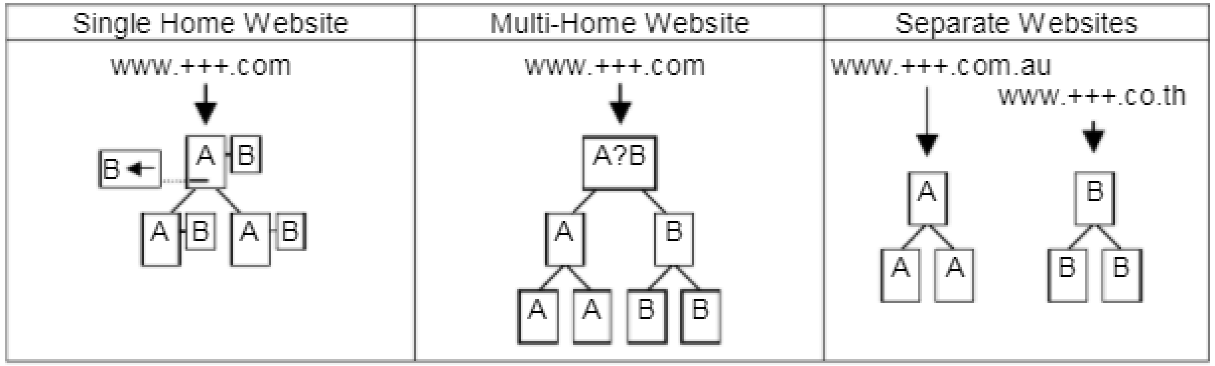
MULTILINGUALITY

- Multilingual - "(of people or groups) able to use more than two languages for communication, or (of a thing) written or spoken in more than two different languages"
- The Bilingualism Versus Multilingualism Dimension
  - For purposes of this task, we will define "Multilingualism" as written or spoken in two or more different languages. Bilingualism or trilingualism are instances of multilingualism.
- W3C: "a “multilingual” web site refers to a web site that uses more than one language."

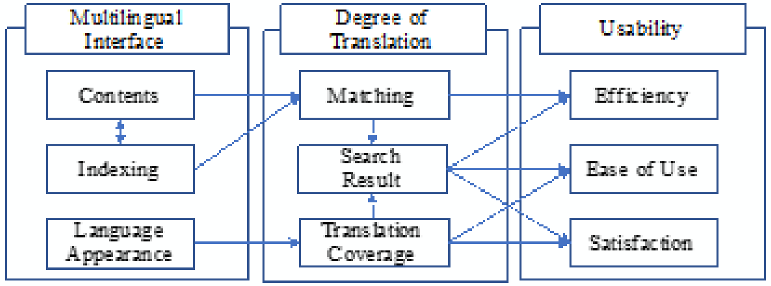


https://www.w3.org/International/questions/qa-international-multilingual
https://en.wikipedia.org/wiki/Multilingualism - picture

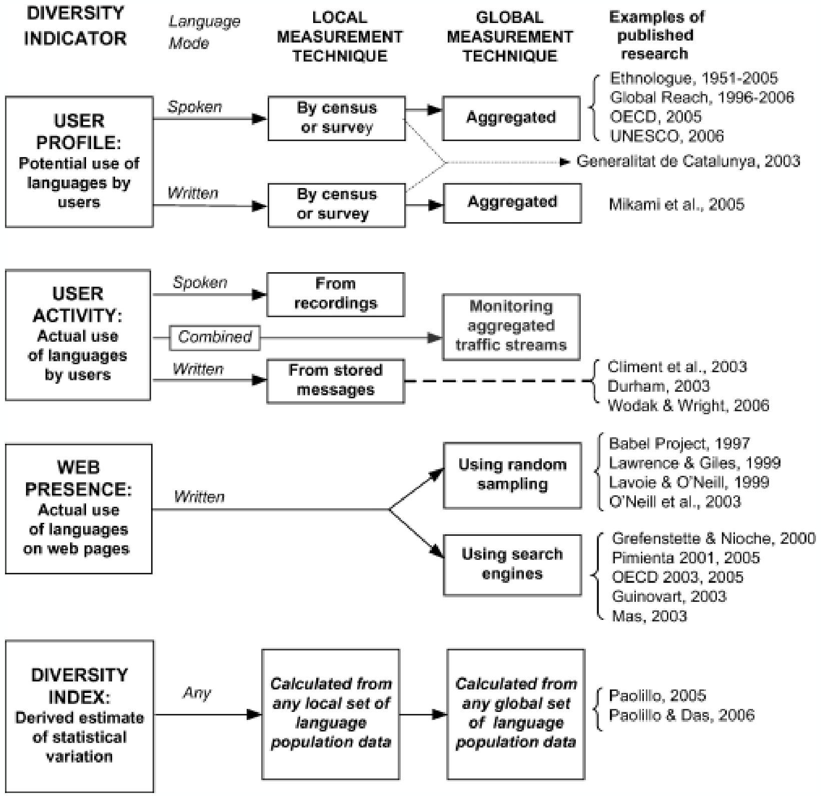
METHODOLOGIES



Three Broad Types of Multilingual Websites



Multilingual Evaluation Guideleni model proposed by Lee and Choi (2019)



A taxonomy of different methodologies used to estimate language diversity on the Internet (Gerrand, 2007)

https://www.researchgate.net/publication/2939572_Multilingual_Website_Usability_Cultural_Context

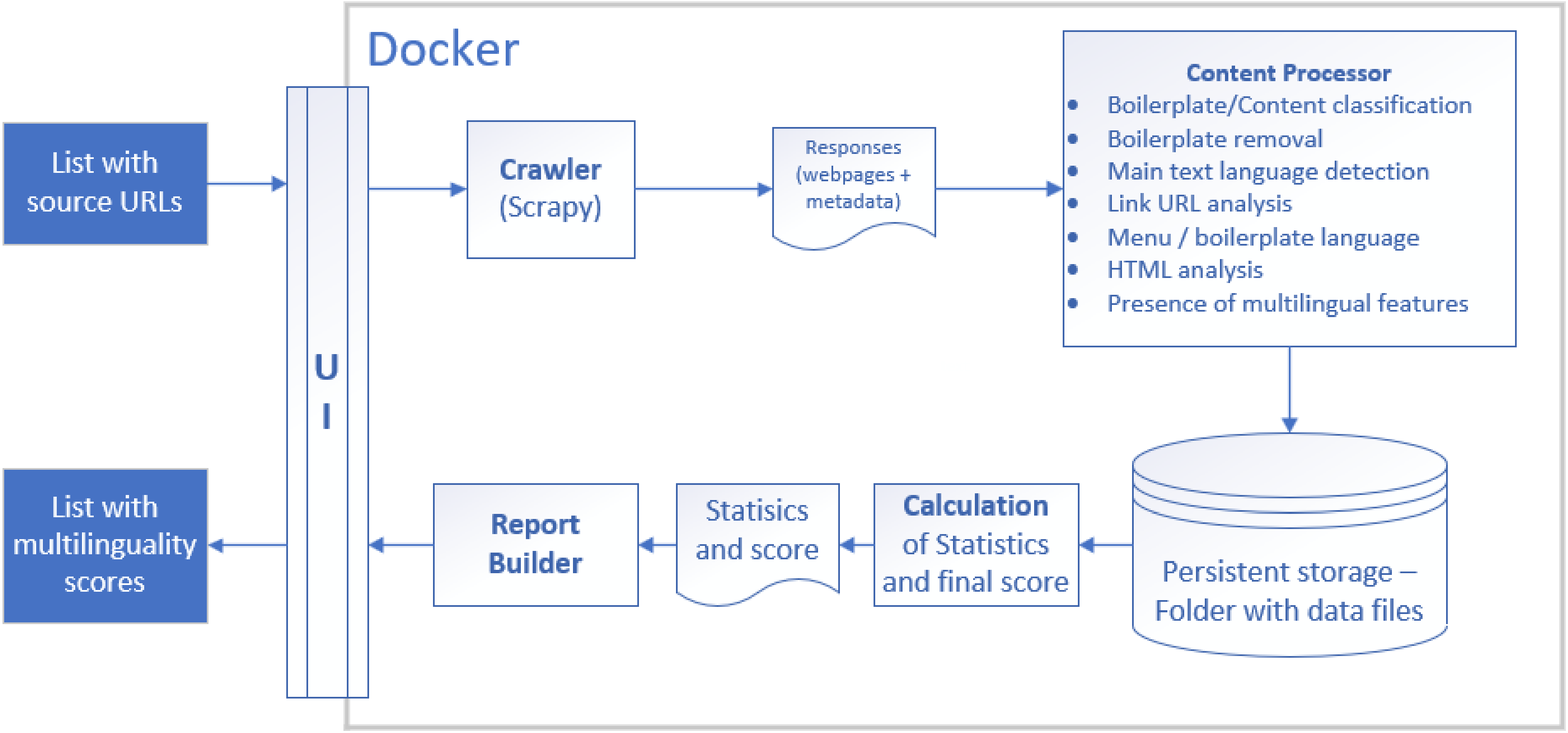
MULTILINGUALITY MEASURE

✓Language coverage

✓Language balance

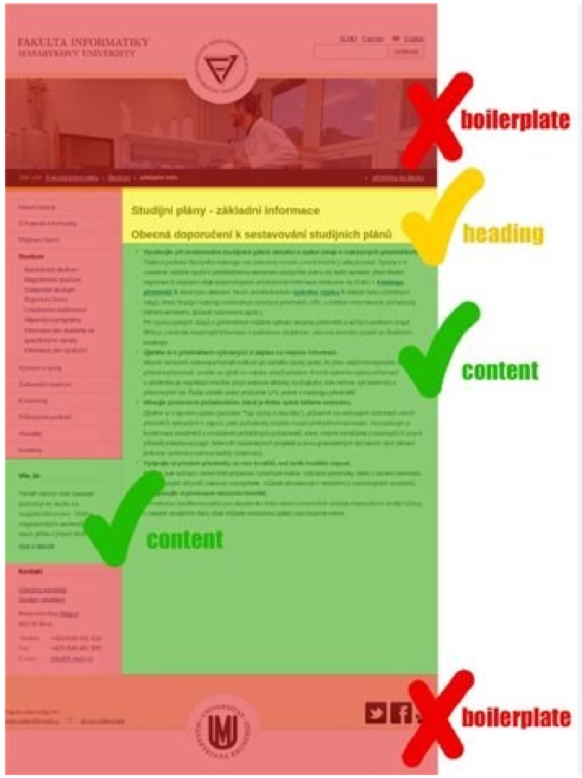
- Linguistic quality
- Technical quality – i18n attributes and others
- Content parallelism and Multilingual functionality (Navigation)

Architecture


Docker
List with source URLs
U I
Crawler (Scrapy)
Responses (webpages + metadata)
Content Processor
Boilerplate/Content classification
Boilerplate removal
Main text language detection
Link URL analysis
Menu / boilerplate language
HTML analysis
Presence of multilingual features
Persistent storage – Folder with data files
Calculation of Statistics and final score
Statisics and score
Report Builder
List with multilinguality scores

COMPONENTS

- Crawler – Scrapy
- Boilerplate removal – jusText
- Language detection – LangDetect
- Scoring – 2 formulas:
  - Normalized language balance = sum(Share1, Share2, …, ShareN)/N
  - Lieberson's diversity index $LDI = 1 - \sum P_i^2$
    - where Pi represents the share of i-th language

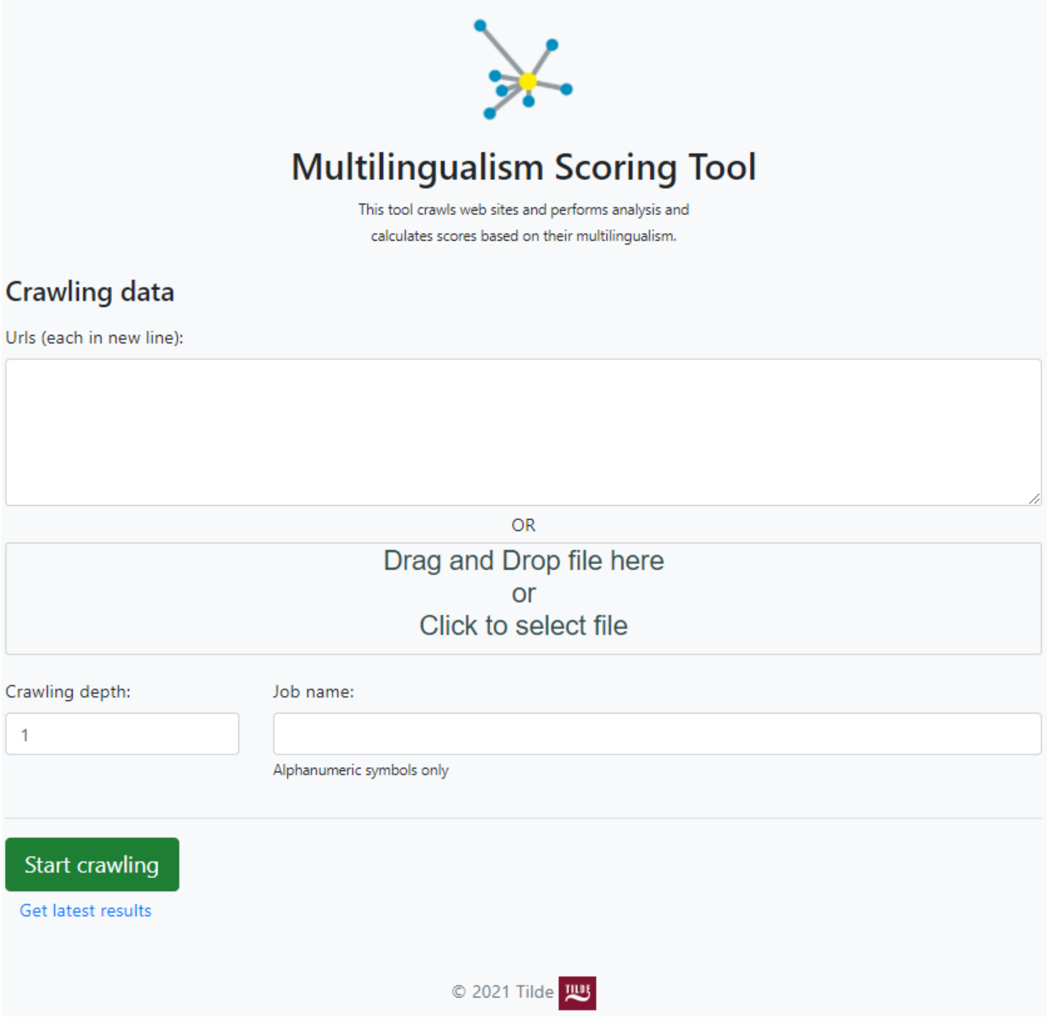


FAKULTA INFORMATIKY
MASARYKOVY UNIVERZITY

Studijní plány - základní informace

Obecná doporučení k sestavování studijních plánů

THE TOOL



Multilingualism Scoring Tool

This tool crawls web sites and performs analysis and
calculates scores based on their multilingualism.

Crawling data

Urls (each in new line):

OR

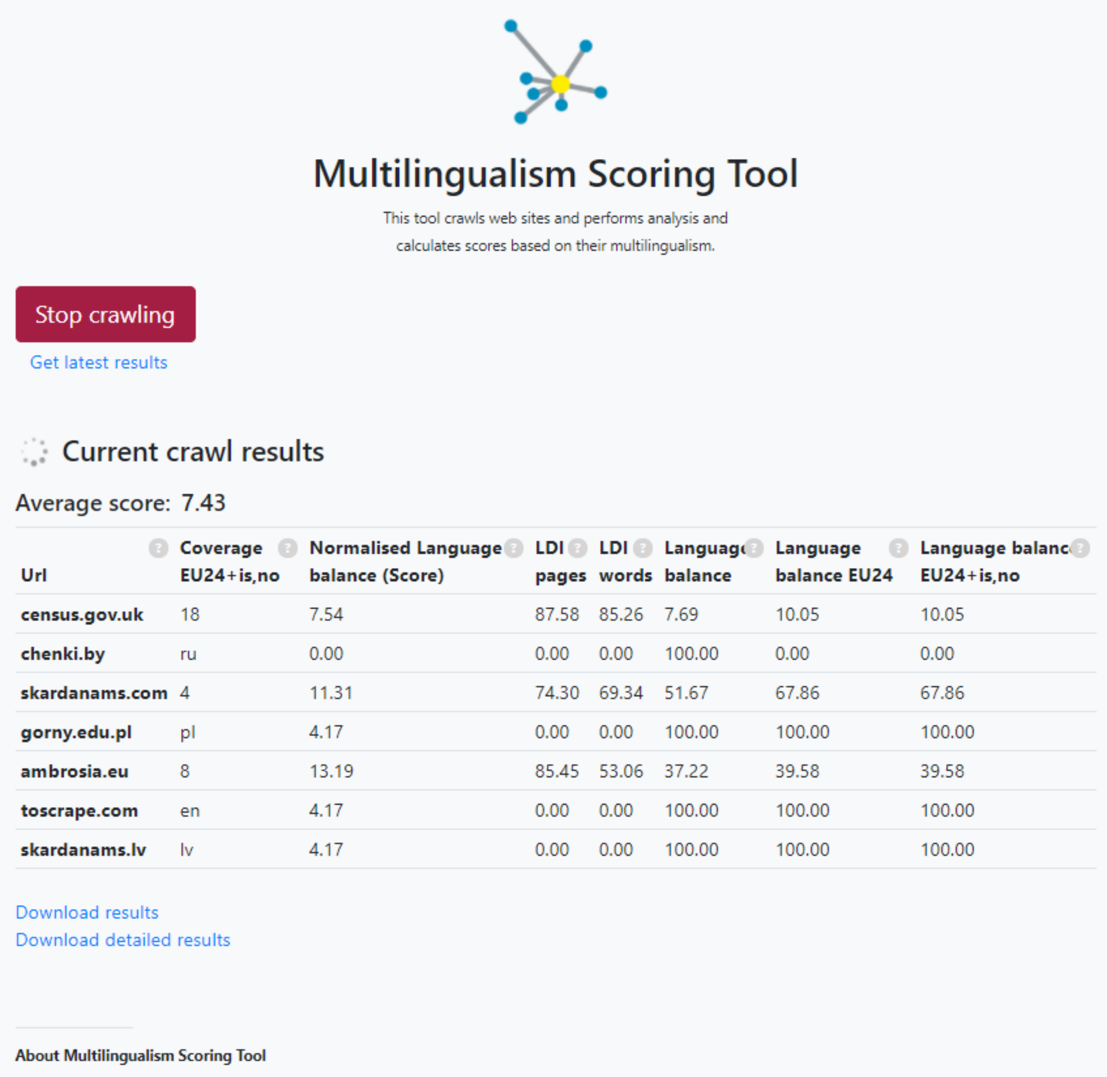
Drag and Drop file here
or
Click to select file

Crawling depth: 1

Job name:

Alphanumeric symbols only

© 2021 Tilde



Multilingualism Scoring Tool

This tool crawls web sites and performs analysis and
calculates scores based on their multilingualism.

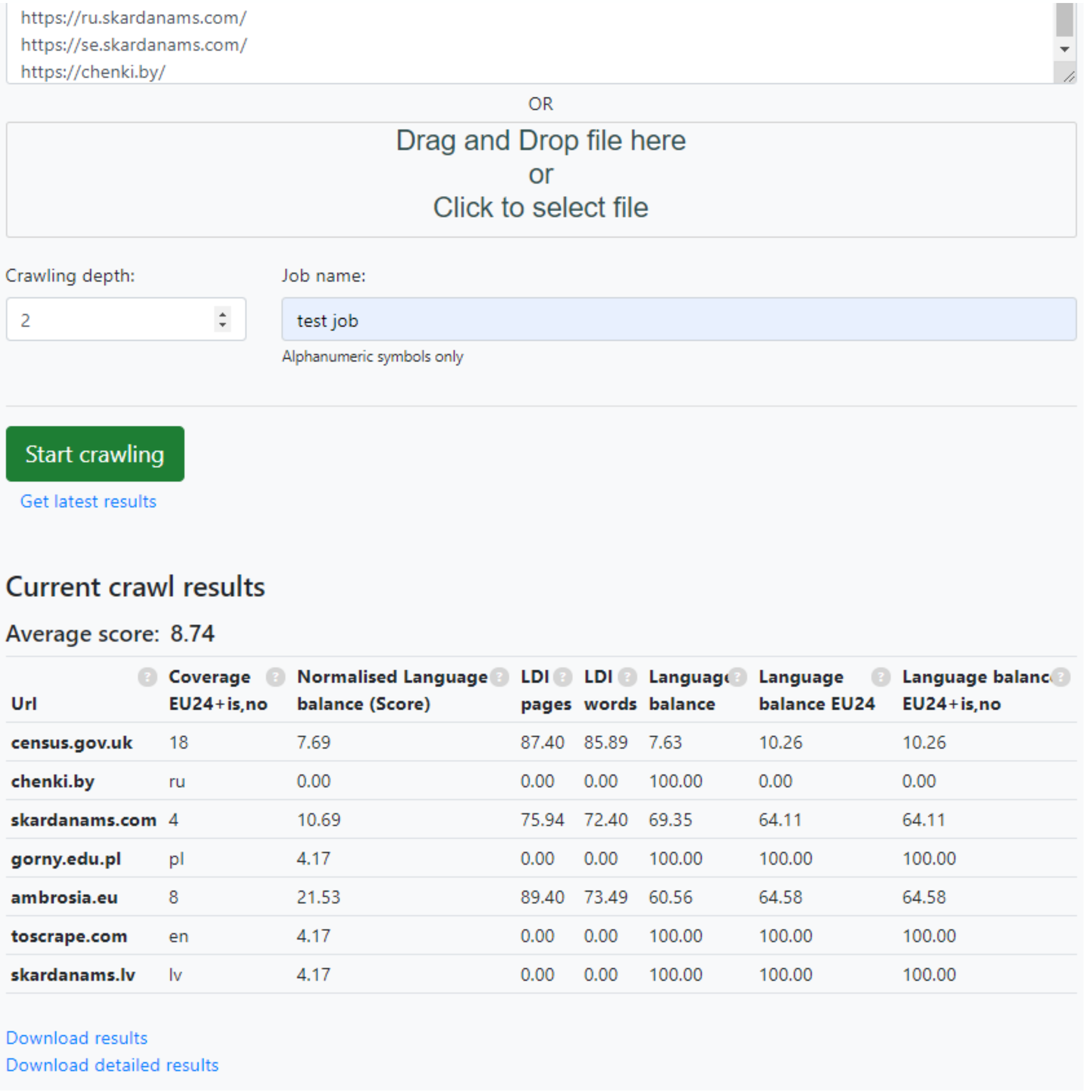
Current crawl results

Average score: 7.43

Url	Coverage EU24+is,no	Normalised Language balance (Score)	LDI pages	LDI words	Languag balance	Language balance EU24	Language balanc EU24+is,no
census.gov.uk	18	7.54	87.58	85.26	7.69	10.05	10.05
chenki.by	ru	0.00	0.00	0.00	100.00	0.00	0.00
skardanams.com	4	11.31	74.30	69.34	51.67	67.86	67.86
gorny.edu.pl	pl	4.17	0.00	0.00	100.00	100.00	100.00
ambrosia.eu	8	13.19	85.45	53.06	37.22	39.58	39.58
toscrape.com	en	4.17	0.00	0.00	100.00	100.00	100.00
skardanams.lv	lv	4.17	0.00	0.00	100.00	100.00	100.00

About Multilingualism Scoring Tool

THE TOOL



https://ru.skardanams.com/
https://se.skardanams.com/
https://chenki.by/

OR

Drag and Drop file here
or
Click to select file

Crawling depth: 2

Job name: test job

Alphanumeric symbols only

Current crawl results

Average score: 8.74

Url	Coverage EU24+is,no	Normalised Language balance (Score)	LDI pages	LDI words	Languag balance	Language balance EU24	Language balanc EU24+is,no
census.gov.uk	18	7.69	87.40	85.89	7.63	10.26	10.26
chenki.by	ru	0.00	0.00	0.00	100.00	0.00	0.00
skardanams.com	4	10.69	75.94	72.40	69.35	64.11	64.11
gorny.edu.pl	pl	4.17	0.00	0.00	100.00	100.00	100.00
ambrosia.eu	8	21.53	89.40	73.49	60.56	64.58	64.58
toscrape.com	en	4.17	0.00	0.00	100.00	100.00	100.00
skardanams.lv	lv	4.17	0.00	0.00	100.00	100.00	100.00

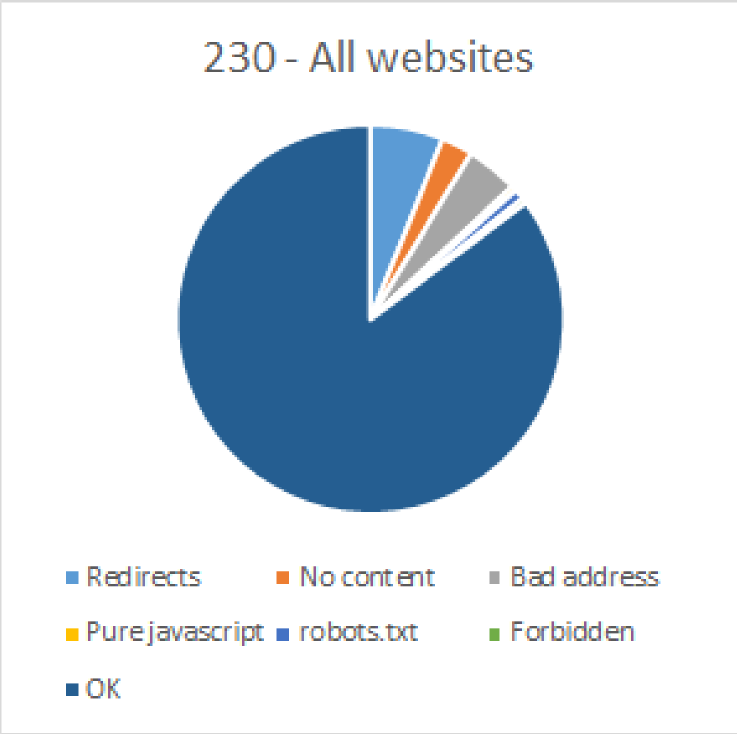
Multilingualism Scoring Tool is available for download either as a code and Docker container on the github: https://github.com/tilde-nlp/Multilingualism-scoring-tool

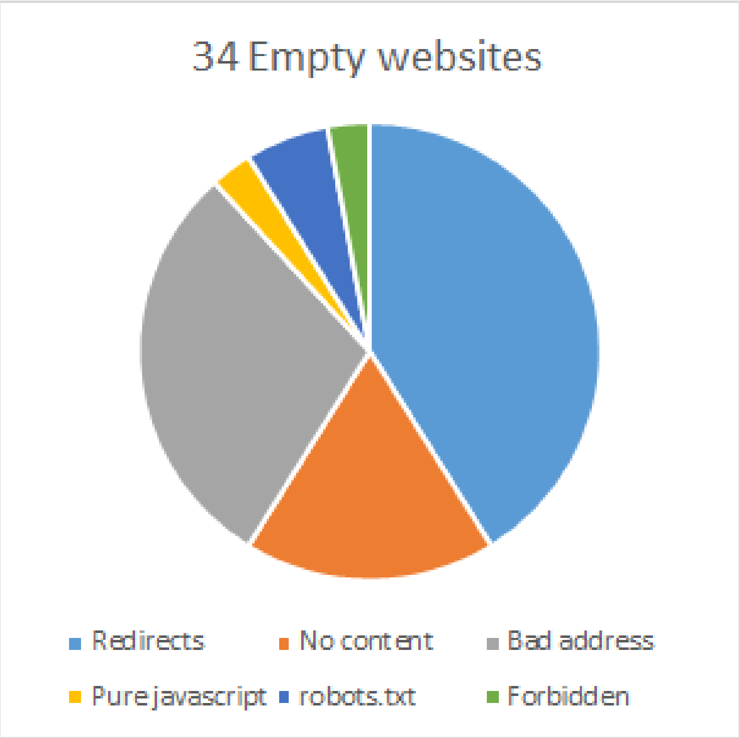
European Language Resource Coordination
Connecting Europe Facility

European Commission

Observations while using tool

- Crawled bottom 230 websites from https://moz.com/top500 with depth 1 in two days
- Found 34 empty results:
  - Redirects 14
  - No content 6
  - Bad address 10
  - Javascript 1
  - Restrictive robots.txt 2
  - Forbidden 1
  - OK 196





Observations while using tool

- Number of prepared requests in each depth (before filtering)
  - 'request_depth_count/0': 14,
  - 'request_depth_count/1': 1000,
  - 'request_depth_count/2': 72004,
  - 'request_depth_count/3': 1259091,
- Stopped after ~110K requests
  - 'exception_count': 17784 detailed breakdown:

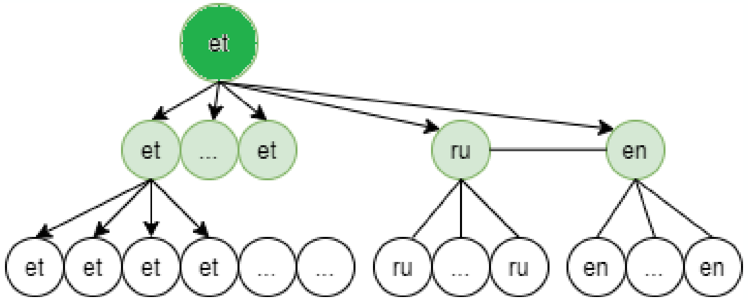
    'ValueError':2, 'IgnoreRequest':12784, 'CancelledError':1, 'ConnectionRefusedError':4929, 'DNSLookupError':54, 'TimeoutError':11, 'ResponseNeverReceived': 3,
  - Responses 92927:

    200 – 73096; 301 – 7534; 302 – 6338; 307 – 598; 404 – 228; 500 – 77, others <15 each

Observations while using tool

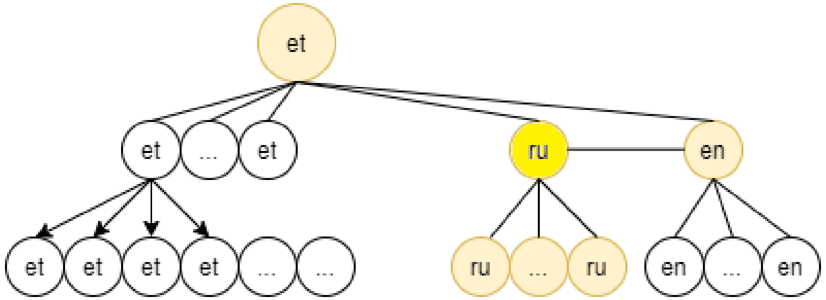
https://president.ee/et/

- 1 Hops: Pages: 45; w/o 8, et 35, ru 1, en 1
- 2 Hops: Pages: 2892; w/o 1274, et 1497, ru 39, en 72, fi 2, lv 1, uk 1, lt 1, el 1, ….
- 3 Hops: Pages: 8272; w/o 3670, et 3164, ru 463, en 910, fi 10, lv 4, uk 5, lt 3, ...
- 4 Hops: Pages: 13179; w/o 6566, et 3938, ru 857, en 1694, fi 24, lv 6, uk 8, lt 9, ...



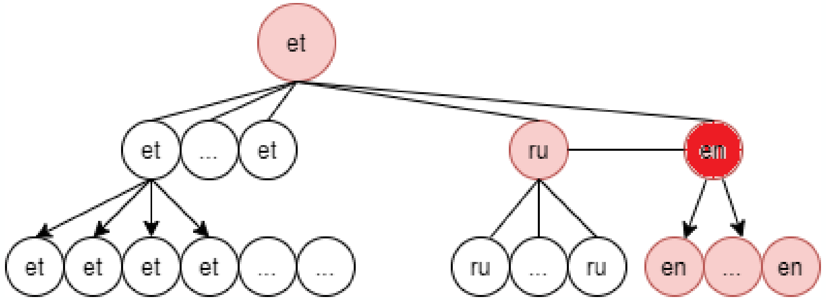
https://president.ee/ru/index.html

- 1 Hops: Pages: 37; w/o 7, ru 27, en 1, et 2
- 2 Hops: Pages: 673; w/o 158, ru 434, en 38, et 42, uk 1
- 3 Hops: Pages: 5373; w/o 2147, et 1540, ru 824, en 812, uk 5, fi 5, lv 3, lt 2, ...
- 4 Hops: Pages: 12031; w/o 5980, et 3367, ru 898, ne 1670, uk 8, fi 19, lv 6, ...



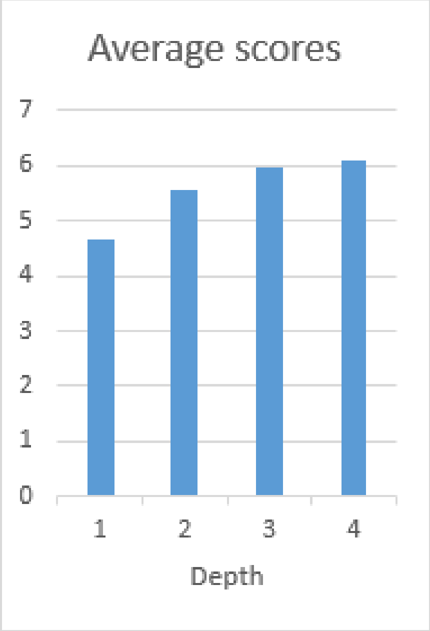
https://president.ee/en/index.html

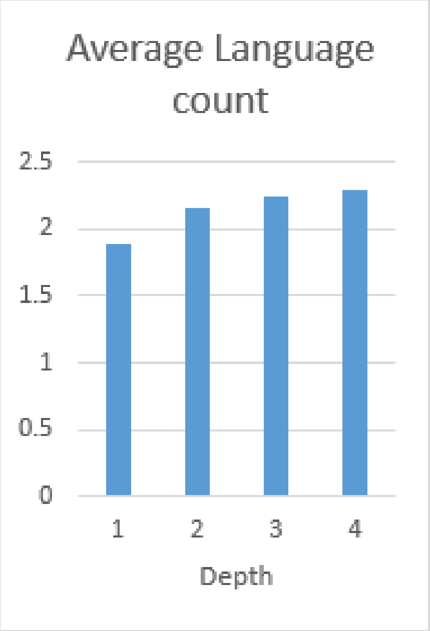
- 1 Hops: Pages: 39; w/o 6, ru 1, en 31, et 1
- 2 Hops: Pages: 1133; w/o 293, ru 28, en 733, et 43, uk 2, fi 3, de 6, sv 2, lv 2, ...
- 3 Hops: Pages: 6162; w/o 2445, et 1625, ru 456, en 1537, fi 14, de 15, sv 4, lv 5, ...
- 4 Hops: Pages: 12471; w/o 6149, et 3421, en 1881, ru 879, fi 28, de 21, sv 5, lv 6, ...

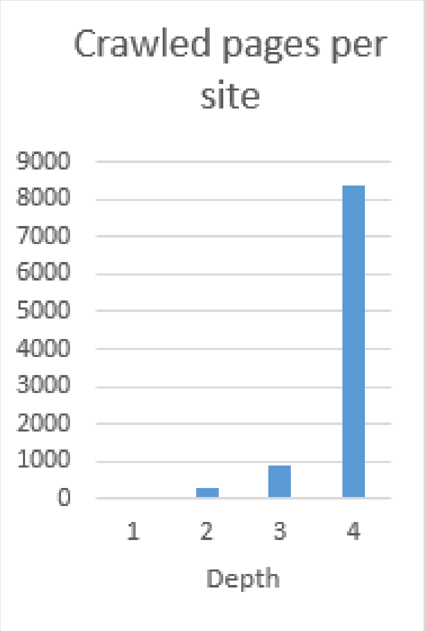


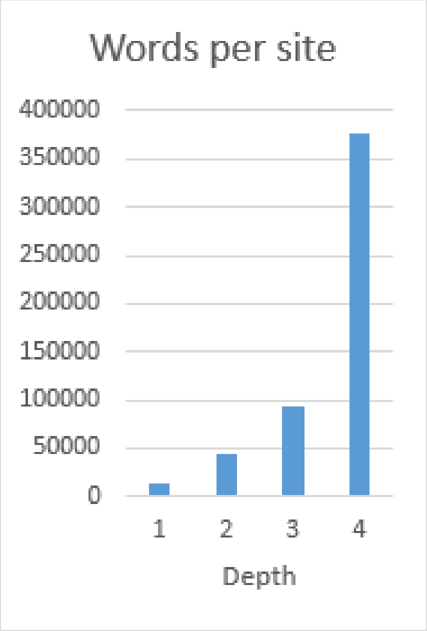
RESULTS

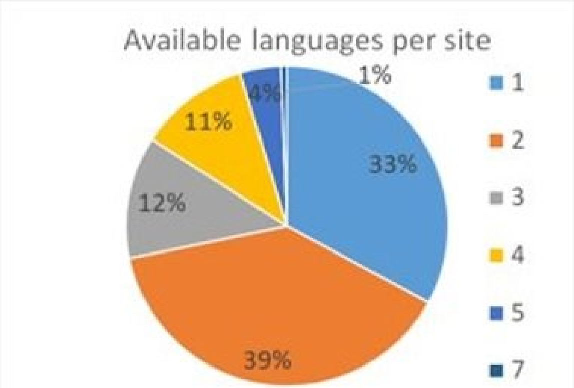
URLs	Crawling depth	Crawl time	Memory usage max	Average score	Average coverage of EU languages	Average number of pages/site	Average number of words/site
198	1	0:50h	160MB	4.65	1.89	33	14516
198	2	12:54h	1030MB	5.57	2.15	262	44592
198	3	48h	1238MB	5.97	2.25	885	93242
198	4	>250h	6672MB	6.08	2.29	8374	376787
600	2	52h	1023MB	5.56	2.02	227	53663











(At depth 3)

CONCLUSIONS

- The tool can crawl a large number of given websites and give some results almost immediately, updating scores as more pages are crawled;

- Time necessary for "polite" crawling increases quickly as we try to exhaustively crawl a website;

- "Multilinguality score" we are currently using represents coverage of translated content in respect to the most represented language on the website, as well as the number of languages the content is translated in;

- European websites currently are not very multilingual – on average content is presented only in 2-3 languages;

European Language Resource Coordination
Connecting Europe Facility

TILDE

European Commission

THANK YOU FOR YOUR ATTENTION!