MINING AND ANALYSING ONE BILLION REQUESTS TO LINGUISTIC SERVICES

Experiences and lessons learned from running a linguistic infrastructure for ten years

Marco Büchler, Thomas Eckart, Greta Franzini, Emily Franzini

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

MOTIVATION

WORTSCHATZ PROJECT

- Collection of corpora in more than 230 languages
- Corpora are collected from e.g. RSS feeds, newspapers and other web content
- Delivers further information such word frequencies, statistically-significant bigrams and co-occurrences from different window sizes

In table 1 it is written ...

Language	*Number of sentences (in M)*	*Language*	*Number of sentences (in M)*
English	1,110	Georgian	30
German	1,023	Bokmål	27
Russian	456	Modern Greek	25
Spanish	244	Lithuanian	20
French	178	Catalan	16
...	...	...	...

Table 1: Text material of the Leipzig Corpora Collection (excerpt)

MOTIVATION FOR THE LEIPZIG LINGUISTIC SERVICES (LLS)

- <2003: individual dumps of the databases were created, partially even with a graphical user interface
- 2004: personnel costs required for this workflow became unsustainable
- 2004/5: Development of a SOAP-based and SOA-oriented infrastructure containing only microservices

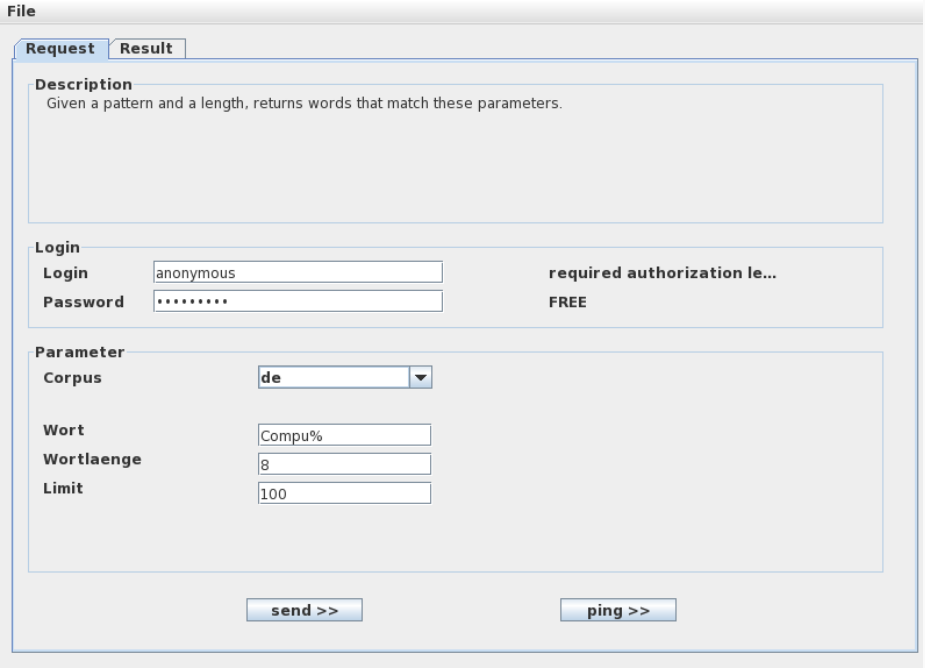
Requirement: a simple but generic architecture that reduces the costs for user responses (email)

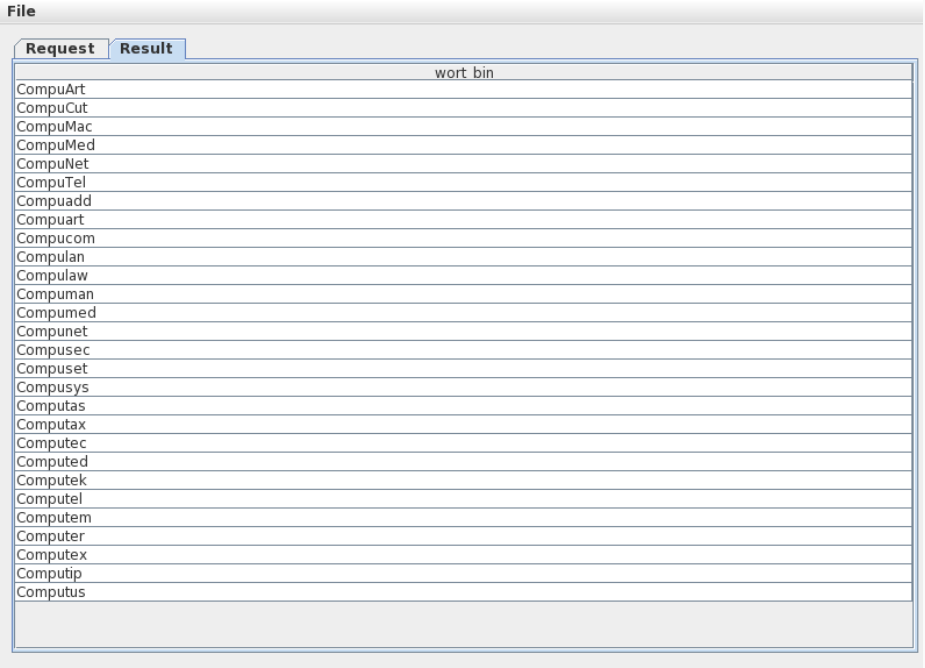
Trade-off: A generic architecture can be reused in different scenarios but tends to have too many parameters and options, while a simple architecture claims usability and guarantees a faster learning curve.

USERS & USER GROUPS

- Research
  - Text profiling and authorship attribution
  - Used as resource for sentiment analysis
- Business
  - Primary interest were services such as Baseform and Synonyms for improving internal search indexes (enterprise search)
  - Usage in portals for weighting words in a word cloud or to display enriching information
- Private
  - A dedicated service was installed upon request to support crossword puzzling
  - Integration in OpenOffice to use e.g. the better Thesaurus-service

AUTOMATICALLY GENERATED GRAPHICAL USER INTERFACES





OPENOFFICE INTEGRATION (EXAMPLE FROM 2005)

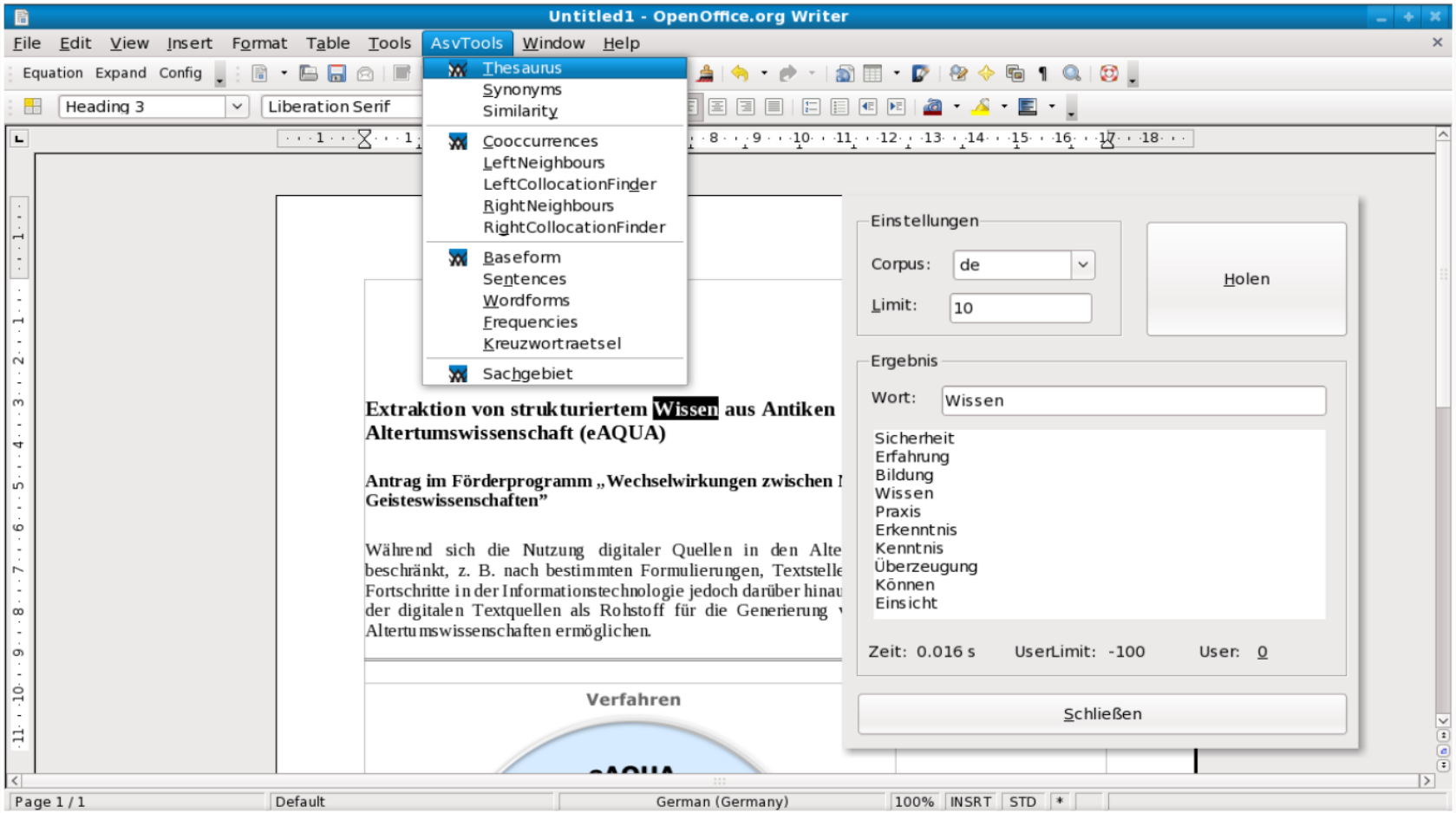


Figure 1: OpenOffice integration of the Leipzig Linguistic Services.

DATA DESCRIPTION

DATA DESCRIPTION OF THE LOG-FILES

Request:

```
2006-09-19T08:43:32+01:00 - anonymous - Baseform - 81.169.187.22 - IN -
0 - execute - Wort=privilegium majus
```

Response:

```
2006-09-19T08:43:32+01:00 - anonymous - Baseform - 81.169.187.22 - OUT -
0 - execute - (0, 0) - 0.03s
```

Remark:

Requests and responses are stored separately in order to be able to detect the number of active requests from log-files.

RESULTS

SERVICE DISTRIBUTION

Service	*Requests*	*Requests (%)*	*Non-empty responses*	*Coverage (%)*	*Input Fields*	*Webservice Type*	*Access level*	*Installation date*
Baseform	624,275,884	64.636%	315,724,185	50.57%	W	MySQLSelect	FREE	04/2005
Category	120,476,452	12.473%	43,276,840	35.92%	W	MySQLSelect	FREE	04/2005
Thesaurus	69,573,648	7.203%	37,151,565	53.39%	W, L	MySQLSelect	FREE	04/2005
Synonyms	60,745,973	6.289%	2,719,544	4.47%	W, L	MySQLSelect	FREE	04/2005
Sentences	60,087,714	6.221%	11,536,172	19.19%	W, L	MySQLSelect	FREE	04/2005
Wordforms	12,671,302	1.311%	4,309,791	34.01%	W, L	MySQLSelect	FREE	04/2005
Frequencies	11,932,213	1.235%	8,095,420	67.84%	W	MySQLSelect	FREE	04/2005
LeftCollocationFinder	1,416,001	0.146%	295,714	20.88%	W, PoS, L	MySQLSelect	FREE	10/2005
RightCollocationFinder	1,379,356	0.142%	235,323	17.06%	W, PoS, L	MySQLSelect	FREE	10/2005
Cooccurrences	1,057,722	0.109%	629,795	59.54%	W, ST, L	MySQLSelect	FREE	04/2005
RightNeighbours	959,560	0.099%	567,870	59.18%	W, L	MySQLSelect	FREE	04/2005
LeftNeighbours	731,449	0.075%	473,600	64.74%	W, L	MySQLSelect	FREE	04/2005
Similarity	467,809	0.048%	308,877	66.02%	W, L	MySQLSelect	FREE	10/2005
CooccurrencesAll	20,852	0.002%	20,848	99.98%	W, ST, L	MySQLSelect	INTERN	05/2009
ExperimentalSynonyms	20,779	0.002%	14,860	71.51%	W, L	MySQLSelect	FREE	12/2009
Crossword puzzling	2,902	$< 0.001\%$	1,306	45.00%	W, WL, L	MySQLSelect	FREE	10/2005
MARSService	616	$< 0.001\%$	616	100.00%	W, L	MARS	INTERN	10/2006
NGrams	564	$< 0.001\%$	149	26.41%	P, L	MySQLSelect	FREE	08/2011
NGramReferences	409	$< 0.001\%$	87	21.27%	P, L	MySQLSelect	FREE	08/2011
Common co-occurrence	55	$< 0.001\%$	43	78.18%	W1, W2, L	MySQLSelect	INTERN	10/2005
TOTAL	965,821,260		425,362,605					

Table II

OVERVIEW OF REQUESTS MADE TO LLS BETWEEN 2006-2014, IN DESCENDING ORDER. THE *Responses* COLUMNS ONLY LIST RESPONSES WHOSE VALUE WAS NOT EMPTY. FOR SPACE REASONS, THE VALUES IN THE *Input Fields* COLUMN ARE ABBREVIATED: *Word* (W.), *Limit* (L.), *Part of Speech pattern* (PoS), *Significance Threshold* (ST), *Word length* (WL) AND *Pattern* (P)

> 30k USERS ON PRECISION VS. RECALL

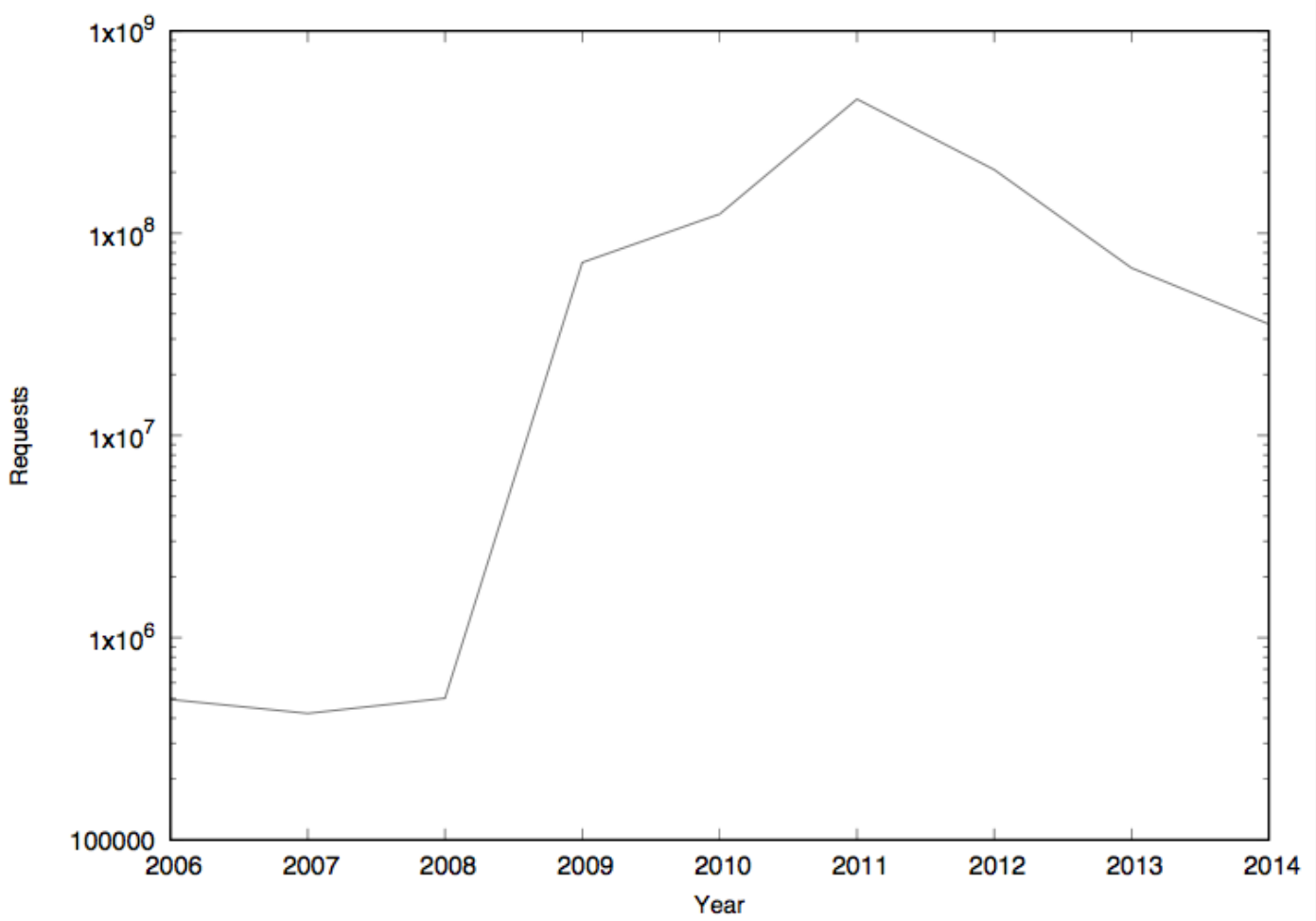
Service	*Requests*	*Requests (%)*	*Non-empty responses*	*Coverage (%)*	*Input Fields*	*Webservice Type*	*Access level*	*Installation date*
Baseform	624,275,884	64.636%	315,724,185	50.57%	W	MySQLSelect	FREE	04/2005
Category	120,476,452	12.473%	43,276,840	35.92%	W	MySQLSelect	FREE	04/2005
Thesaurus	69,573,648	7.203%	37,151,565	53.39%	W, L	MySQLSelect	FREE	04/2005
Synonyms	60,745,973	6.289%	2,719,544	4.47%	W, L	MySQLSelect	FREE	04/2005
Sentences	60,087,714	6.221%	11,536,172	19.19%	W, L	MySQLSelect	FREE	04/2005
Wordforms	12,671,302	1.311%	4,309,791	34.01%	W, L	MySQLSelect	FREE	04/2005
Frequencies	11,932,213	1.235%	8,095,420	67.84%	W	MySQLSelect	FREE	04/2005
LeftCollocationFinder	1,416,001	0.146%	295,714	20.88%	W, PoS, L	MySQLSelect	FREE	10/2005
RightCollocationFinder	1,379,356	0.142%	235,323	17.06%	W, PoS, L	MySQLSelect	FREE	10/2005
Cooccurrences	1,057,722	0.109%	629,795	59.54%	W, ST, L	MySQLSelect	FREE	04/2005
RightNeighbours	959,560	0.099%	567,870	59.18%	W, L	MySQLSelect	FREE	04/2005
LeftNeighbours	731,449	0.075%	473,600	64.74%	W, L	MySQLSelect	FREE	04/2005
Similarity	467,809	0.048%	308,877	66.02%	W, L	MySQLSelect	FREE	10/2005
CooccurrencesAll	20,852	0.002%	20,848	99.98%	W, ST, L	MySQLSelect	INTERN	05/2009
ExperimentalSynonyms	20,779	0.002%	14,860	71.51%	W, L	MySQLSelect	FREE	12/2009
Crossword puzzling	2,902	< 0.001%	1,306	45.00%	W, WL, L	MySQLSelect	FREE	10/2005
MARSService	616	< 0.001%	616	100.00%	W, L	MARS	INTERN	10/2006
NGrams	564	< 0.001%	149	26.41%	P, L	MySQLSelect	FREE	08/2011
NGramReferences	409	< 0.001%	87	21.27%	P, L	MySQLSelect	FREE	08/2011
Common co-occurrence	55	< 0.001%	43	78.18%	W1, W2, L	MySQLSelect	INTERN	10/2005
TOTAL	965,821,260		425,362,605					

Table II

OVERVIEW OF REQUESTS MADE TO LLS BETWEEN 2006-2014, IN DESCENDING ORDER. THE *Responses* COLUMNS ONLY LIST RESPONSES WHOSE VALUE WAS NOT EMPTY. FOR SPACE REASONS, THE VALUES IN THE *Input Fields* COLUMN ARE ABBREVIATED: *Word* (W.), *Limit* (L.), *Part of Speech pattern* (PoS), *Significance Threshold* (ST), *Word length* (WL) AND *Pattern* (P)

Lessons learned: Users prefer precision over recall.

NUMBER OF REQUESTS PER YEAR BETWEEN 2006 AND 2014



Lessons learned: Don't change the settings of a running system!

WHAT DID USERS ALSO SEND?

Cleanliness of requests:

Rule	*Matched requests (in % of all)*
Broken encoding	66, 869, 667 (6.920%)
Query too short	2, 978, 216 (0.310%)
URLs, HTML code, email addresses, etc.	189, 895 (0.019%)
Query too long (more than 200 characters)	69, 799 (0.007%)

Table 2: Applied rules for "cleanliness" of queries (excerpt)

Lessons learned: At least 71 million request (7.4%) are noise from crawled and badly extracted web content.

HOW DID USERS COMBINE REQUESTS TO FORM CHAINS?

Detected and useful service chains:

Rank	*Service chain*	*Percentage*
1	Baseform Frequencies	67.11%
2	Baseform Synonyms Sentences	26.32%
3	Synonym Sentences	3.00%
4	Baseform Synonyms	1.01%
5	Baseform Frequencies Synonyms	0.97%
6	Baseform Thesaurus	0.68%
7	Baseform Frequencies Category	0.24%
8	Baseform Category	0.24%
9	Frequencies Baseform Frequencies	0.23%
10	Thesaurus Similarity	0.20%

Table 3: List of top-ten most frequently discovered service chains

Six chains, represented by the ranks 2, 4, 5, 6, 7 and 8, following the `Baseform * [Synonym|Thesaurus|Category]*` pattern.

MINED SERVICE CHAINS BY PRAGMATIC USAGE

However, chains such as:

```
Baseform Synonyms Sentences Baseform Synonyms Sentences
```

were more critical as they doubled one of the core chains.

This discovery can be explained with the following example:

```
If I had had enough flour, I would have made more
                                    brownies.
```

Lessons learned: Automatic installation of aggregated chains is not feasible. However, the discovery helps to identify candidates followed by human judgement.

FURTHER RESULTS IN THE PAPER

- Suggestions for load balancing based on user requests
- Influence of multi-word units on the results
- Corpus-building, corpus representativeness and corpus balancing
- Interoperability issues of the SOAP protocol in different programming languages
- Results of benchmarks for SOAP- and REST-based web-services
- For REST-based services: Comparison of standoff- vs. inline-markup

CONTACT

Speaker

Marco Büchler, Thomas Eckart, Greta Franzini, Emily Franzini.

Visit us

http://www.etrap.eu

contact@etrap.eu

teckart@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG

eTRAP

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

SPONSORED BY THE

Federal Ministry of Education and Research

ARCHITECTURE OF THE LLS

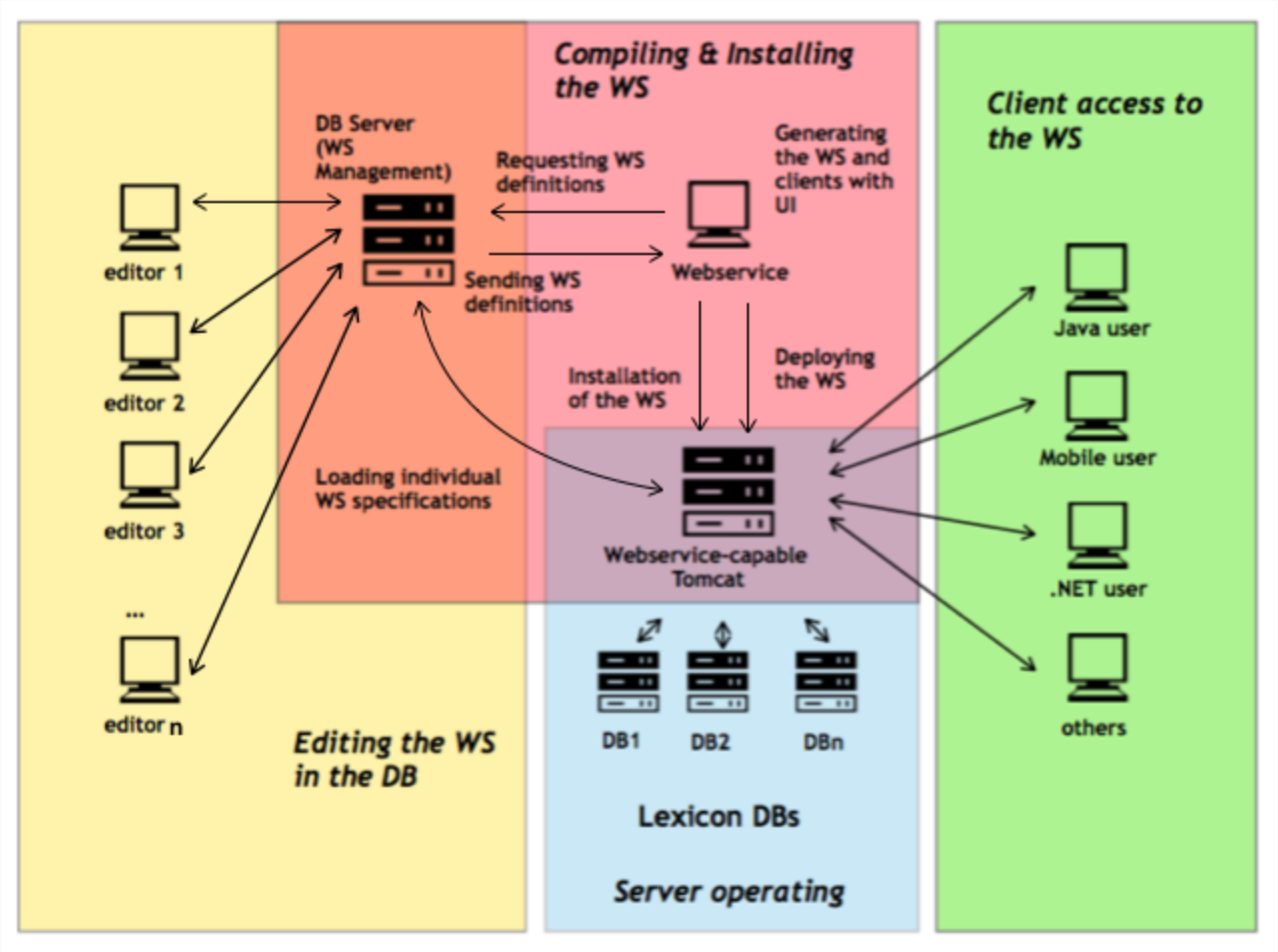


Figure 2: Four workflow modes with separation of concern: editing (yellow); managing, compiling and deploying (red); hosting and operating (blue); using the LLS infrastructure (green).

GEOGRAPHICAL DISTRIBUTION OF THE LLS

Country	*Requests*	*Percentage*
Germany (DE)	921, 184, 562	99.29%
Ireland (IE)	2, 003, 348	0.22%
Swiss (CH)	1, 957, 431	0.21%
Austria (AT)	1, 347, 703	0.13%
Hungary (HU)	302, 966	0.03%
Poland (PL)	212, 357	0.02%
Japan (JP)	184, 408	0.02%
Romania (RO)	90, 140	0.01%
China (CN)	90, 125	0.01%
France (FR)	82, 969	$<$ 0.01%

Table 4: Top-ten list of requests by country for the years 2006 - 2014

LICENCE