Diversity through localization:

How ccTLDs enhance linguistic diversity online

Emily Taylor

Table of Contents

1. Overview

The world’s websites speak English. There are more than 6,000 languages[1] spoken in the world, but online English is stubbornly popular, the language of 54% of web pages. Major world languages such as Arabic and Hindi, with more than a billion speakers offline, are barely represented in web pages[2].

As part of CENTR’s 20th anniversary celebrations, CENTR requested Oxford Information Labs to test the hypothesis that ccTLDs support local languages. Thanks to the cooperation of CENTR members, Oxford Information Labs has been given unprecedented access to zone files and language analysis for 16.4 million geographic domain names.

In a study of 16.4 million domains managed by CENTR members, on average 76% of web content reflected local languages.

CENTR members and associate members collectively manage 80% of all registered country code domain names worldwide[3]. Many were set up during the 1990s, prior to the commercialization of domain name markets. The majority were set up to reflect the ethos of the early internet, as expressed in RFC 1591, that ccTLD managers are ‘trustees for the delegated domain and have a duty to serve the community’[4]. That sense of service to local internet communities naturally includes operating and supporting local languages.

In this study, we found that on average, 76% of web content associated with each TLD reflects the languages spoken in the relevant country or territory. The English language accounted for 19% of web content, and other languages 4%.

For those TLDs in the study which also offer internationalized domain names or IDNs (ie domains with accents, diacritics and non-Latin scripts), local languages represented a higher proportion of web content (84%), and English a lower proportion (9%), consistent with the finding that IDNs help to enhance linguistic diversity in cyberspace[5].

In keeping with RFC 1591, ccTLD managers are ‘trustees for the delegated domain and have a duty to serve the community’.

[1] https://en.unesco.org/news/unesco-launches-website-international-year-indigenous-languages-iyil2019

[2] See W3Techs ‘Usage of content languages for websites, 2019 https://w3techs.com/technologies/overview/content_language/all

[3] CENTRStats Global TLD Report, April 2019 https://stats.centr.org/stats/global

[4] https://www.rfc-editor.org/rfc/rfc1591.txt

[5] https://idnworldreport.eu/

2. Methodology

Most country code Top Level Domains do not make their zone files publicly available. Researchers interested in measuring online linguistic diversity rarely have access to ccTLD raw data.

Oxford Information Labs' research team has developed its own methodology for automated web content language analysis over many years as part of the annual EURid UNESCO World Report on Internationalized Domain Names. Training our algorithms on the publicly available generic Top Level Domains (gTLD) has facilitated the analysis of large data sets, with sources in differing formats. For this study, the research team further refined its methodology for identifying single page and parking sites.

As part of CENTR's 20th anniversary papers series, a call went out to the CENTR community asking for their collaboration. The response was positive. This study represents a language analysis of 10 TLDs, comprising 16.4 million domain names. The data set represents 16% of the domains managed by CENTR members and associate members.

The data was shared within the period of December 2018 to May 2019.

Through automated data analysis, the research team performed the following tasks:

- Identify domains with active services.
- Automated language analysis for all active domains in the data set.
- Identify low-quality content (single page and suspected parking pages).
- Compare language analysis before and after elimination of low-quality content.
- Identify internationalized domain names (IDNs) in the data sample and compare all the above results with the full data set.

The detailed methodology is set out in the appendix.

3. Which TLDs were analyzed?

TLD	Country or territory	Principal languages[6]	# domains in sample	Notes
.cat	Catalonia	Catalan	0.1 m	Not a ccTLD, a gTLD supporting the Catalan community
.ch	Switzerland	German, French, Italian	2.1 m	
.dk	Denmark	Danish	1.3 m	
.nl	Netherlands	Dutch	4.5 m	SIDN, the .nl ccTLD registry, shared high-level results of its own language analysis with the research team.
.nu	Nuie	Swedish	0.5 m	Operated by the Swedish ccTLD registry
.pt	Portugal	Portuguese	0.3 m	Total .pt zone is 1.1m[7]; DNS.pt shared a sample of the zone.
.ru / .рф	Russian Federation	Russian	5.8m	Combined zones of .ru and .рф TLDs
.se	Sweden	Swedish	1.4 m	
.sk	Slovakia	Slovak	0.4 m	
Total domains in study			**16.4 m**	

Table 1: summary of top level domains in study

[6] Source: Ethnologue

[7] Data correct to 22 May 2019, source https://www.dns.pt/pt/estatisticas/

4. Results

4.1.1 Domains with active services

To be capable of supporting any language content, a domain name must have active services. On average, the percentage of domains with active services (nameservers or email) per TLD in our analysis is 80%. Results vary widely: the highest percentage of active domains was found in .sk (Slovakia) with 91%, and the lowest in .nu (Nuie) with 44%.

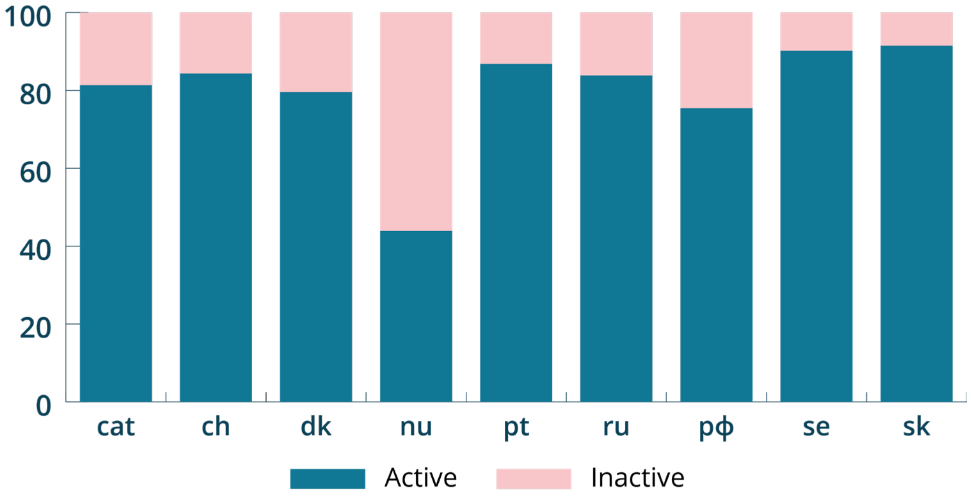


Figure 1: domains with active services

For IDNs the rate of active services was on average 72% with the lowest being IDNs at the second level under .cat at 36% and the highest .pt at 93%.

The results for some TLDs may not be representative of the entire zone, as some TLD operators shared a sub-set of the zone, comprising domains with active services, or eliminated domains scheduled for deletion as part of the standard renewal cycle.

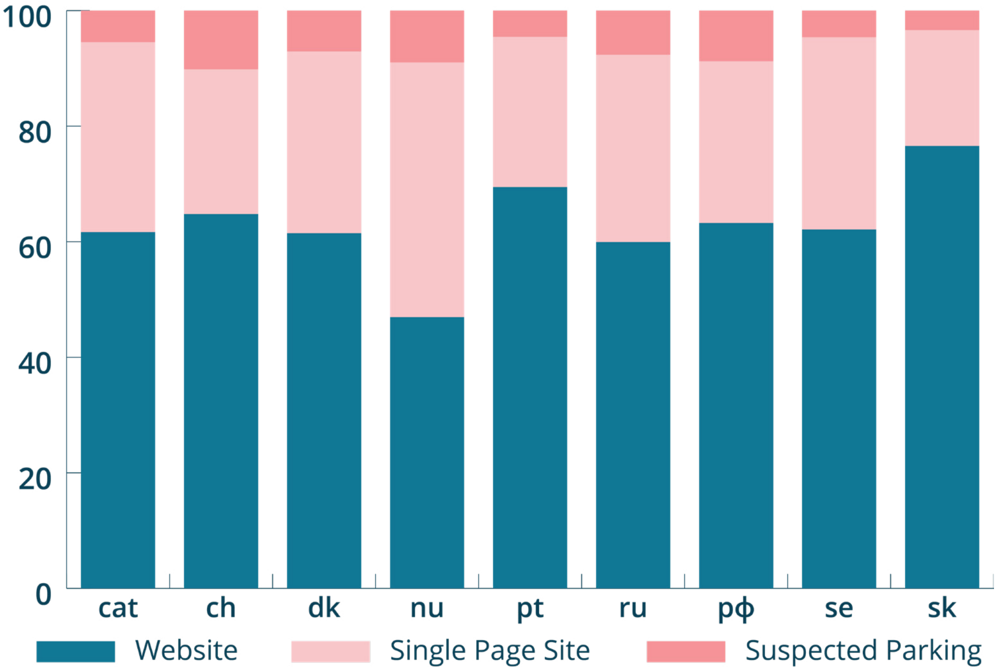
The purpose of this analysis is not to pass any judgement on the usage rates within a particular TLD, but to eliminate from further analysis those domains categorized as inactive.

4.1.2 Use of domain names: identify low quality content

Registering a domain name is often one of the first steps taken as a new business or venture is set up. It may take months or years for an active site to be built. During this period the registrar or hosting provider may put up a single page or parking site. Single page and parking sites are usually pro-forma with identical or similar wording. Parking sites are sometimes associated with pay-per-click activity to monetize web traffic.

The research team identified as low-quality content, and eliminated from further study, the following:

- Single page sites
- Suspected parking sites (see detailed methodology)



On average, low-quality content (parking and single pages) accounted for 37% of domains in the study.

Figure 2: domain name usage by TLD

On average the combined percentage of low-quality content was 37%. The highest rate was under .nu (53%) and the lowest, .sk (23%).

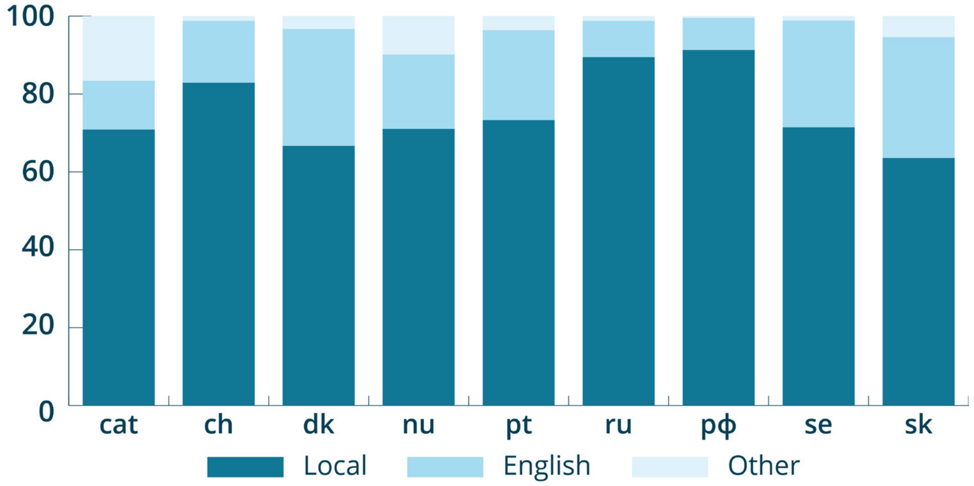
The IDNs within the data set, both at top and second level, had a higher rate of low-quality content (43%). The highest rate was for IDNs under .nu (58%) and the lowest .pt (2%). The .sk ccTLD does not support IDNs.

4.1.3 Language analysis

Excluding domains with low-quality content, the research team undertook an automated analysis of the language of web content associated with the remaining domain names in the data sample, in accordance with the methodology.

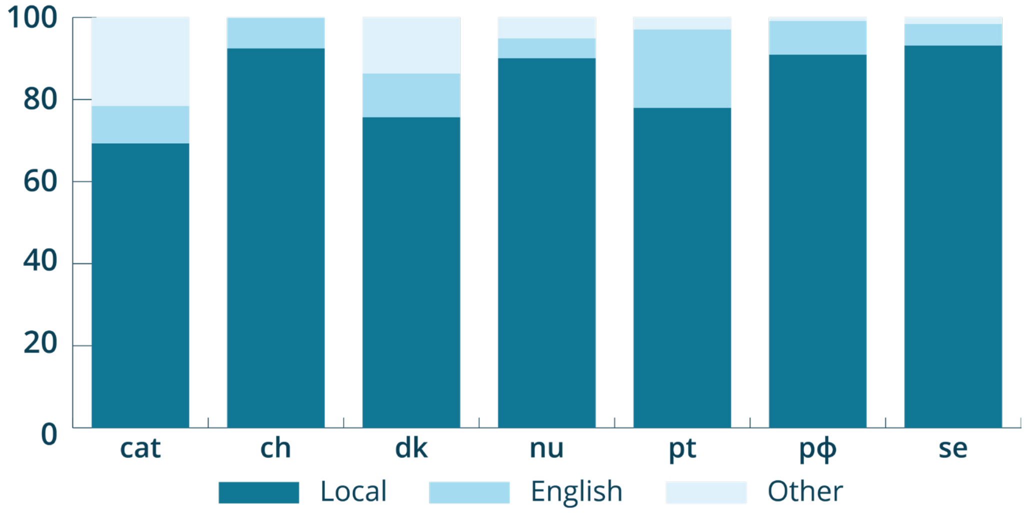
4.1.3.1 Local language usage

Without exception, the principal languages spoken in the country or territory comprised at least 64% of the content in the zone, and the average rate per TLD was 76%.



On average, local language content comprised 76% of the websites in each TLD.

Figure 3: language analysis per TLD (excluding low-quality content)



Local language sites comprise 84% on average for the IDNs in the study

Figure 4: language analysis of IDNs per TLD (excluding low-quality content)

Where the TLD also supported IDNs, whether at the top or second level, the average rate of local language content rises to 84%, with the rates of 90% and above under .ch, .nu, .se and .рф. The lowest rate of local language in IDNs was under .cat at 69%.

4.1.3.2 The Dutch ccTLD

SIDN, the manager of the ccTLD for the Netherlands, .nl, shared with the research team a summary of its own language analysis of the Dutch zone.

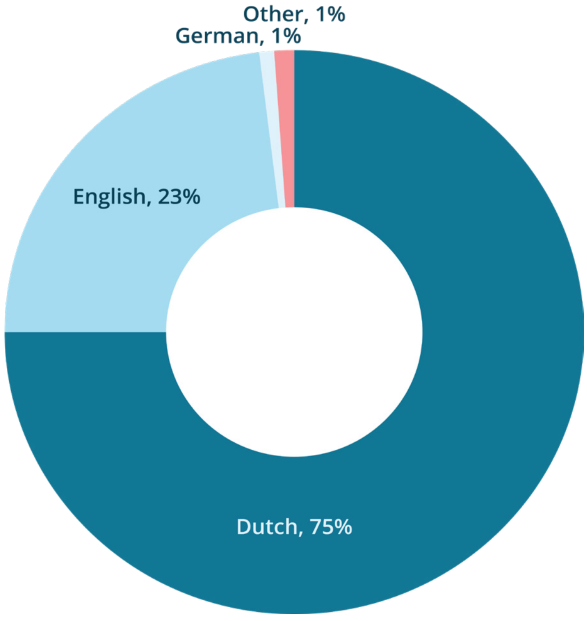


Figure 5: language analysis of .nl TLD (courtesy of SIDN)

The results for .nl are consistent with the other TLDs in the study, with the local language Dutch accounting for 75% of the content in the zone. It is not known whether the SIDN methodology excluded single page or suspected parking sites.

4.1.3.3 Rates of English language – the impact of low-quality content

When the research team compared our language analysis before and after the exclusion of low-quality content, there was a significant decrease in the percentage of English language sites across all TLDs.

	Entire zone		IDNs only	
TLD	Before	After	Before	After
.cat	50%	12%	58%	9%
.ch	37%	16%	30%	7%
.dk	48%	30%	41%	11%
.nu	50%	19%	42%	5%
.pt	44%	23%	19%	19%
.ru	26%	9%		
.рф (IDN)	16%	8%	16%	8%
.se	43%	27%	39%	8%
.sk	41%	31%		
OVERALL AVERAGE	**39%**	**19%**	**35%**	**9%**

Table 2: English content before and after elimination of low-quality content

On average, the rate of English language per TLD before elimination of low-quality content was 39%, with the highest rate being found in .cat and .nu (50%) and the lowest under .ru (26%). After the elimination of low-quality content, the average rate of English language per TLD drops to 19%, with the largest decrease being seen in .cat (from 50% to 12%) and the smallest decrease in .sk (41% to 31%).

English is more likely to be the language of low-quality content, such as parking and single page sites.

Among the IDNs (both top and second level), the average rate of English content per TLD before elimination of low-quality content was 35%, with the highest rate being found in .cat (58%) and the lowest .pt (19%). After the elimination of low-quality content, the average rate of English language per TLD drops to 9%, with the largest decrease being seen in .cat (from 58% to 9%) and the smallest decrease in .pt, where the rate is unchanged.

4.1.3.4 Presence of indigenous languages

2019 is UNESCO's International Year of Indigenous languages[8], which identifies 2,680 languages as being in danger of extinction. Europe has several languages on UNESCO's list of endangered languages, such as Corsican, Galician, Irish, Welsh and Basque. The web environment favours English and major languages, with endangered languages being even more rare in web content than they are in the offline world.

We checked to see whether the TLDs in our sample had indigenous or endangered languages represented in the language of web content. We were interested to see whether the Sámi language (which has approximately 30,000 speakers across Scandinavia[9]) was present in the .se or .nu zones, but Sámi is not yet supported by Google translate. Sadly, spot-checks of other endangered languages flagged in our analysis turned out to be errors in the automated translation tool.

There was little evidence of indigenous languages among ccTLDs in the study.

[8] UNESCO international year of indigenous languages 2019 https://en.iyil2019.org/

[9] UNESCO Atlas of the World's Languages in Danger, http://www.unesco.org/languages-atlas/index.php

5. Findings and conclusions

This study's findings indicate that country and regional TLDs boost the presence of local languages online and show lower levels of English language than is found in the domain name sector worldwide.

Country and regional TLDs boost the presence of local languages online.

The pattern of language usage is not random, but matches the languages spoken in the country or territory represented by the TLD. So, Slovak, which accounts for 0.4% of the world's websites[10], is the language of 64% (91,000+) of .sk domains. Likewise, the percentage of 'other' languages in each zone is low (typically less than 5%), indicating that internet users view each ccTLD as reflecting the geographic country or territory and its languages.

Table 3 summarizes the findings across each of the TLDs in the study.

TLD	Local language of country or territory	Active usage	Parking+	#1 language	#2 language	#3 language *	Other languages
.cat	Catalan	81%	38%	Catalan	Spanish	English	4%
.cat IDNs		36%	53%				0%
.ch	German, French, Italian, Romansch	84%	35%	German	English	French	1%
.ch IDNs		77%	43%				1%
.dk	Danish	80%	39%	Danish	English	Swedish	3%
.dk IDNs		52%	52%				3%
.nl	Dutch	No information		Dutch	English	German	1%
.nu	Swedish Danish Dutch	44%	53%	Swedish	English	Dutch	5%
.nu IDNs		86%	58%			Danish	3%
.pt	Portuguese	87%	31%	Portuguese	English	Spanish	2%
.pt IDNs		93%	2%				2%
.ru	Russian	84%	40%	Russian	English	Bulgarian	<2%
.рф		75%	37%				<1%
.se	Swedish	90%	38%	Swedish	English	German	<2%
.se IDNs		86%	55%	Swedish	English	Romanian	<1%
.sk	Slovak	91%	23%	Slovak	English	Czech	3%

+ Parking includes single page and suspected parking sites (see section 4)

*the #3 language is included in 'other' languages in charts under section 4.

Table 3: summary of findings

[10] W3Techs, 2019 https://w3techs.com/technologies/overview/content_language/all

The top three languages of each zone, excluding English, reflect the principal languages spoken in the country or territory represented by the TLD. The second and third most occurring languages also relate to the geographical region and language families. So, Spanish was found in the .cat and .pt zones; Czech in the .sk zone; and Swedish in the .dk zone. The one anomaly is the presence of Romanian in IDN second-level .se domains, but the percentage is very low (<1%).

English was strongly present in all TLDs in our study, but in all cases was far below the global average of 54%[11].

Language analysis of the IDNs in the study showed lower levels of English language and higher levels of local languages than the larger data sets of ASCII domains.

The rates of English language were far below the global average.

After the elimination of low-quality content, ie single page and suspected parking sites, the percentage of English language sites fell across every TLD in the study. This leads to the conclusion that English is more likely to be the language of low-quality web content.

While ccTLDs show consistent alignment with the principal languages spoken in the relevant country or territory, the presence of indigenous or minority languages is weak. Despite the global nature of the web, the languages online do not reflect those of the real world. In this context, the role of country and regional TLDs is all the more essential in supporting online linguistic diversity.

The contribution of ccTLDs to linguistic diversity online is seldom appreciated, because few publish data relating to language usage, and the majority do not make their zones freely available for research purposes. Therefore, few studies focus on language use within ccTLDs. Through the cooperation of the CENTR community, Oxford Information Labs was given rare access to the raw data or language analysis of more than 16.4 million domains to explore the research question of the extent to which ccTLD and geographic TLD enhance local language usage online.

[11] ibid

Appendix Full methodology

Data gathering

Through the CENTR network, we invited CENTR members and associates to share their zone file or language analysis with us.

The research team had available raw data from nine TLDs, .ch, .dk, .se and .nu, .pt, .ru and .рф, .sk. In addition, we were able to query the zone for the new gTLD for the Catalan community, .cat, through ICANN's CZDS File. In total, the research team had access to TLD zones comprising a total of 12 million records. SIDN shared a summary of its own language analysis for 4.5 .nl domains. In all, the study represents a language analysis of more than 16.4 million domain names, or 16% of the domains managed by CENTR members and associate members.

There were some differences in the format in which registries shared their zones. Some, such as .se and .nu, make their entire zone publicly available. Others, such as .pt shared only those domains that were actively resolving. Therefore, the total number of domains in the study does not always correlate exactly to the total number of domains under management for each TLD. The methodology followed by SIDN for the .nl zone may differ from that of the Oxford Information Labs research team.

The data was shared within the period of December 2018 to May 2019.

Data analysis

Identify active domains

The research team ran each domain in the zones through the following tests:

- Check for nameserver and MX records. A [or AAAA] nameservers are present in the domain name record, MX records are present, check for www record.
- Eliminate records which do not have any of the above present from further analysis, and mark as 'inactive'

From these tests, we are able to report on the active usage of domains in each TLD. Results may be affected if the registry shared only active domains with the research team. This would tend to skew the results for those TLDs in favour of active usage.

For the remaining set of domains, the research team performed the following checks:

- Check whether domain name has active web content; store status code and content of first page.
- Identify redirects and collect redirect domain
- Check for parking hints: number of internal links (identifying single page sites); content siblings (ie identical content); redirection siblings where many domains are redirecting to the same domain name; number of content words is greater than 50; language of web content is Latin ('lorem ipsum'… placeholder text).

Run all active domains through language analysis

Automated language analysis runs through the following steps:

- Using stored content, identify the 10 most frequently occurring keywords ignoring stop words
- Run stored keywords through automated translation tools as individual words
- To eliminate potential errors in the automated translation, additional runs are performed on subsets of content data are performed using full phrase translation.

Re-run language analysis excluding single page and suspected parking sites

The automated language analysis is repeated, following the removal from the data sample of domains with sites identified as single page and suspected parking.

Identify 'local languages' for each TLD (see table 1)

IDN analysis

Identify domain names in the data set beginning xn-- as internationalized domain names (IDNs). Repeat steps 2.2.1 to 2.2.3

Caveats

The research team takes care to ensure the quality of the language analysis, but errors do occur. The automated translation tool employed (Google Translate) is usually accurate, but occasionally makes errors. Where languages share common words (e.g. Swedish and Danish; Slovak and Czech) the automated translation tool can wrongly identify one for the other. The tool also struggles with rare languages, such as Galician, Corsican, Irish, and Welsh – incorrectly assigning an endangered language.

We are not convinced that all the domains assigned as 'Spanish' in the .pt zone are in fact Spanish. Manual checking of a handful showed that several were Portuguese. Likewise, under the .sk zone, several of the domains identified as having Czech language proved to be in Slovak on manual checking. The research team re-ran phrases from those data sets (rather than individual keywords) through the automated translation tool but errors are likely to persist.

Council of European National Top-Level Domain Registries

CENTR is the association of European country code top-level domain (ccTLD) registries, such as .de for Germany or .si for Slovenia. CENTR currently counts 54 full and 9 associate members – together, they are responsible for over 80% of all registered domain names worldwide. The objectives of CENTR are to promote and participate in the development of high standards and best practices among ccTLD registries. Full membership is open to organisations, corporate bodies or individuals that operate a country code top level domain registry.

This paper is part of a series of articles covering industry research, historical data analysis and the future of technologies such as digital IDs, published over the course of 2019 to mark CENTR's 20th Anniversary. These publications do not necessarily present the views of CENTR or of the CENTR community.

CENTR wishes to thank and acknowledge the organisations which have so generously contributed to the efforts of its 20th Anniversary:

Platinum sponsor

Gold sponsor

Silver sponsor

CENTR vzw/asbl
Belliardstraat 20 (6th floor)
1040 Brussels, Belgium
Tel: +32 2 627 5550
Fax: +32 2 627 5559
secretariat@centr.org
www.centr.org

To keep up-to-date with CENTR activities and reports, follow us on Twitter, Facebook or LinkedIn