NZRS
an InternetNZ company

Technical Research Report

Introduction

This is the November 2016 Technical Research Report, setting out the research projects being undertaken by NZRS. This is the fourth issue of this report. An updated version of this document will be provided at each Council meeting.

Scope and output of technical research

Technical research aims to expand the frontiers of our knowledge about the Internet within NZ and make that new knowledge openly available to all. Projects are drawn from the wide range of topics within this broad ambit.

One of the earliest considerations is what data is available as data analysis is the cornerstone of research activity. This explains the inevitable heavy emphasis on .nz research in the projects listed below as the data is readily available after a number of years building a data collection and analysis infrastructure for .nz.

Research projects are initiated with an idea of what might be achieved, how that might be used and in what forms the output might be delivered. The identification of potential uses looks beyond research team to consider how other researchers might build on that knowledge and how that knowledge might be commercialised, both within and without NZRS, to aid the growth of the NZ economy.

As with all true research though, there is no guarantee that this is what will be achieved or that the project will not change radically over time and it is not uncommon for a project to change focus or even name during its lifetime.

Wherever possible the outputs of technical research projects will be open knowledge, open code published on our GitHub repository and open data published on our Internet Data Portal (IDP), all under a Creative Commons license. The limitations on this are: a) to respect the privacy inherent in any data used; b) to preserve the security of the Internet; and c) to comply with .nz policies and procedures.

Projects

Title	NZ IP Topology Map	Status	In Progress
Description	Mapping the internal structure of the Internet in New Zealand. This project uses the RIPE Atlas probes to do active measurement and discovery of Internet Topology.		
Potential uses	There are a number of outstanding questions about the structure of the NZ Internet whose answers can drive useful policy debate. For example, are their routes where traffic between one NZ site and another NZ site is forced to sub-optimally 'trombone' out of the country and back again because of the way that some providers interconnect?		
Planned outputs	*Form*	*Done*	*Details*
	Web site	☒	Website at http://ip.topology.net.nz updated with new version.

	Open data	☒	Resulting network representation made available via the project's website.
	Open code	☒	Code available in NZRS GitHub account.
Presented	Proof of Concept presented at First NZIRF. Working version presented at Second NZIRF. Introduced as project seeking involvement at the RIPE 72 Hackathon. Presented a Spain-centric version at the Spain Network Operators Group in October 2016. Presented the methodology at the RIPE 73 meeting in Madrid in the same month.		
Collaborators	No active collaborators at the moment.		
Progress	Needs work automating the execution to make it a regular collection. Make the raw data available via IDP.		

Title	NZ BGP Topology Map	Status	On Hold
Description	Mapping the structure of the Internet in New Zealand using publicly available data sources. Uses BGP feeds from RouteViews, RIPE and data made available by the Internet Exchanges.		
Potential Uses	Understand how the structure of the Internet in New Zealand changes with the pass of time, how different IXs gain/loose peers, etc.		
Planned outputs	*Form*	*Done*	*Details*
	Report	☐	
	Web site	☐	http://bgp.topology.net.nz A new faster version will be made available soon.
	Open data	☒	Collected data made available via IDP
	Open code	☒	Code available in NZRS Github account
Presented	Presented at First NZIRF and previous version at NZNOG 2014.		
Collaborators	None.		
Progress	A new version was written to allow using publicly available APIs, and to store the collected data in IDP. A better visualization, easier to use has been produced and will be deployed to production soon.		

Title	ANZSIC classification of the register	Status	In progress
Description	Using web content from each domain web page, and a set of hand curated domain names mapped to an economic activity code (ANZSIC), train a machine learning model and be able to classify every domain in the register. This allow us to augment our understanding of the register. This work now has been extended to classify non for profit organization using the New Zealand Standard Classification of Non-Profit Organizations (NZSCNPO) from StatsNZ.		
Potential uses	The data could be provided to registrars for their Domains under management (DUMs) in the registrar portal and so help them understand their customers		

	better. The same data could also be made available to registrants through a new product or service.		
Planned outputs	*Form*	*Done*	*Details*
	Open data	□	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants
	Open code	□	Will publish code on GitHub
Presented	Concept presented at 2015 Registrar conference.		
Collaborators	Initial data used for the training was bought from two companies one of which, WhoIsWhere, participated in the first round of analysis of the results.		
Progress	Using a strict mapping from domain to activity code, 50% (+/- 1%) of the testing data was mapped correctly. If using fuzzy matching (any of the top 3 most probable categories), this value increases up to 78% +/- 1% accuracy. Future steps include a better text collection from the webpages to support JavaScript, and better input data clean-up. The non-profit classification is currently at 95% accuracy using strict matching.		

Title	Domain Retention Prediction	Status	In Progress
Description	Project to generate a probabilistic model that will tell us: • Which elements of a registration are best predictors of their likelihood to be stay in the register • Probability of a domain to be stay in the register in the future, and by extension, determine the forward value of a domain in the register		
Potential uses	Can be provided to registrars for their DUMs to enable them to understand their customers better. This work may also allow NZRS to produce a better income forecasting model.		
Planned outputs	*Form*	*Done*	*Details*
	Report	☒	A couple of blog posts are published in NZRS's blog.
	Open code	□	Will publish code on GitHub.
Presented	Concept presented at Registrar Conferences in 2014 and 2015.		
Collaborators	Some of the insights obtained in this work has been shared and discussed with staff at .CA. People from .IE (Ireland) and Netherlands (.NL) are following up this work closely.		
Progress	A rigorous creation forecast model has been produced and published. An analysis and model of domain survivability is available using open data and open code. The following task will be Machine Learning to identify the most relevant elements in a domain affecting cancellations.		

Title	Registrant Classification	Status	On Hold

Description	Machine Learning classifier to determine if a registrant is a person or an organization based on the registrant name.		
Potential uses	Augment our understanding of the register, as this information is not available at registration. Likely this will feed into other research projects rather than have much utility on its own.		
Planned outputs	*Form*	*Done*	*Details*
	Open data	□	Will consider aggregated and anonymised data on IDP.
	Open code	□	Will be published on GitHub.
Presented	None.		
Collaborators	None.		
Progress	Code refactored to improve accuracy and quality of documentation, achieving 96% accuracy. Currently 60.6% of the domains are registered by Organizations, 39.4% by Individuals.		

Title	Domain Popularity Algorithm	Status	In Progress
Description	Algorithm using DNS data to determine if a domain name is more popular than others.		
Potential uses	Can be shared with registrars to help them understand their customers better. Can be used for interesting information about the .nz namespace for the general public in press releases and the like. Can be used to develop new products/services that allow registrants to see how their actions affect their domain name popularity.		
Planned outputs	*Form*	*Done*	*Details*
	Report	□	
	Web site	☒	Some selected data sets are publicly visualized at http://domain-rank.nzrs.net.nz/popular.html and http://domain-rank.nzrs.net.nz/bank.html
	Open data	□	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants
	Open code	□	Will be published on GitHub.
Presented	Presented as Proof of Concept at DNS-OARC 22 in Amsterdam. Presented at the CENTR Jamboree in Brussels in May 2016. Follow up work presented at the DNS-OARC 25 in Dallas, October 2016.		
Collaborators	Seeking collaboration within the CENTR group, as suggested by the CENTR R&D Chair.		
Progress	A review of the algorithm has been done and we are now working with a different approach that produces better results. A sample of DNS traffic from one of your offshore providers will be used to test for bias. Working in identifying well known sources of traffic to treat that traffic in a different way.		

	Google Analytics figures from 4 different domain names to be used to test correctness.

<table>
<tr><th>Title</th><th colspan="2">DGA detection algorithm</th><th>Status</th><th>On Hold</th></tr>
<tr><td>Description</td><td colspan="4">We gave our summer intern relatively free rein to explore our DNS data set and what he came up with is the bones of an algorithm to automatically detect traffic generated by botnets using DGAs (Domain Generation Algorithms) using DNS traffic.</td></tr>
<tr><td>Potential uses</td><td colspan="4">Can be used for early detection of infected hosts. Can be used to assess the overall health of .nz. Can be used to assess the likelihood that a new registration is nefarious in intent.</td></tr>
<tr><td rowspan="3">Planned outputs</td><td><i>Form</i></td><td><i>Done</i></td><td colspan="2"><i>Details</i></td></tr>
<tr><td>Report</td><td>□</td><td colspan="2"></td></tr>
<tr><td>Open code</td><td>□</td><td colspan="2">Will be published on GitHub.</td></tr>
<tr><td>Presented</td><td colspan="4">The concept was presented at the New Zealand Internet Research Forum 2015.</td></tr>
<tr><td>Collaborators</td><td colspan="4">Details have been exchanged with SIDN Labs as they are working in similar ideas.</td></tr>
<tr><td>Progress</td><td colspan="4">The proof of concept needs to be tested at a larger scale, possibly using a different language.</td></tr>
</table>

<table>
<tr><th>Title</th><th colspan="2">Register word decomposition</th><th>Status</th><th>On Hold</th></tr>
<tr><td>Description</td><td colspan="4">Decompose every domain in the register into their word components (aucklandaccountants.org.nz into “Auckland accountants”).</td></tr>
<tr><td>Potential uses</td><td colspan="4">Largely as a building block for other potential projects, such as identifying prevalence of geographic terms (and thereby understanding potential for a new geographic TLD), detecting trending words in registrations and identifying use of Te Reo.</td></tr>
<tr><td rowspan="4">Planned outputs</td><td><i>Form</i></td><td><i>Done</i></td><td colspan="2"><i>Details</i></td></tr>
<tr><td>Report</td><td>□</td><td colspan="2"></td></tr>
<tr><td>Open data</td><td>□</td><td colspan="2">Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants</td></tr>
<tr><td>Open code</td><td>□</td><td colspan="2">Will be published on GitHub.</td></tr>
<tr><td>Presented</td><td colspan="4">None.</td></tr>
<tr><td>Collaborators</td><td colspan="4">None.</td></tr>
<tr><td>Progress</td><td colspan="4">Using a curated list of 2000 domains, and using the LINZ Gazetteer data as input, the classifier achieves an 88% accuracy. Requires a valid Te Reo Māori corpus to increase accuracy.</td></tr>
</table>

Title	Full web scan of .nz	Status	On Hold

Description	Capture web content published under .nz domains to feed the ANZSIC classification project. Investigate tools to do a deeper gathering of content.		
Potential uses	Multiple possible uses including a general report on the state of the .nz web space; information for registrars on their DUMs; information for registrants as part of a new product or service; and as an input into another research projects.		
Planned outputs	*Form*	*Done*	*Details*
	Report	□	
	Open data	□	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants
	Open code	□	Will be published on GitHub.
Presented	None		
Collaborators	We have discussed this project with the National Library who have a contract for a web scan using similar technology and are looking at techniques to mine that data once gathered.		
Progress	A first working version is available and being used for ad-hoc shallow web scans. A second version is available to identify the cases where sites require Javascript to render content, to fetch them using a different tool. A Proof of Concept for the deep scan is available using Hadoop, Heritrix and HBase.		

Title	Zone Scan V2	Status	Not started
Description	The regular zone scan is using code that is no longer maintained. The replacement version allows faster scanning, and easier ways to run custom collections. This work aims to investigate, test and eventually replace the engine used by the zone scan.		
Potential uses	NZRS development team already working on building outputs from v1 into the registrar portal to provide registrars with information on their domains with a view to improving quality. Data could also be provided to registrants in a new product or service.		
Planned outputs	*Form*	*Done*	*Details*
	Open data	□	Will publish aggregated and anonymised data on IDP.
	Open code	□	Will be published on GitHub.
Presented	None		
Collaborators	IIS, the .SE register are collaborators as authors of the engine currently in use, and developers of the replacement.		
Progress	Not Started		

Title	DNS statistics publication using IDP	Status	Not started
Description	Make data from the DNS traffic for .nz available using the Internet Data Portal		

Potential uses	Researchers and Policy makers are always interested in data. DNS data is rich and vast, and can be useful to observe the uptake of new technologies. Making data from the DNS traffic for our ccTLD available in an open format can help the community to answer some questions, like the uptake of IPv6 or DNSSEC. We aim to make some of that data available on a regular basis.		
Planned outputs	*Form*	*Done*	*Details*
	Report	□	
	Open data	□	Will publish aggregated and anonymised data on IDP.
	Open code	□	Will be published on GitHub.
Presented	None.		
Collaborators	SIDN is publishing some interesting counters from their DNS data, using a platform powered by Hadoop, inspired by the work we did with Hadoop		
Progress	Not started		

Title	Digital Journey publication using IDP	Status	Finished
Description	Make data collected from the Digital Journey website about businesses self-assessment of their use of digital technologies available using the Internet Data Portal		
Potential Uses	Data collection started in 2014, and could provide a consistent view on how businesses have evolved their preparedness around digital technologies.		
Planned outputs	*Form*	*Done*	*Details*
	Report	□	
	Web site	□	
	Open data	□	Available in IDP https://idp.nz/Users-and-Use/Digital-Journey/sp2s-ukz9
	Open code	□	
Presented	None.		
Collaborators	MBIE as drivers of the initiative, Firebrand as developers and maintainers of the website.		
Progress	Initial upload of data completed with data from March 2014 to July 2016. Monthly updates scheduled.		

Glossary

Botnet	A network of compromised PCs that are remotely controlled, generally for criminal purposes.
DGA	Domain Generation Algorithm. A technique used by botnets to automatically generate domains names that they can register and use for their command and control servers.

DNS-OARC	The main membership organisation focused on DNS research.
GitHub	The main web site used in our industry for sharing code.
IDP	Our Internet Data Portal at https://idp.nz
NZIRF	New Zealand Internet Research Forum. Organised by InternetNZ.
Hadoop	Big Data Platform