NZRS an InternetNZ company

Technical Research Report

Introduction

This is the July 2017 issue of the Technical Research Report, setting out the research projects being undertaken by NZRS and their status. This document is updated quarterly.

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. Key considerations in this endeavour are:

- Data availability – research projects require data for analysis and therefore data availability is critical to progressing the project. NZRS has built a significant data collection and analysis infrastructure for .nz and therefore many projects lever off this infrastructure.

- New knowledge, uses and outputs – new projects need to consider why the research is being undertaken and how it might be used (including commercialisation) by NZRS or others.

- No guarantees – for all novel research there is risk of not achieving the outcome or that objectives adapt to ongoing research outcomes. This is inherent in research.

- Openness - wherever possible the outputs of technical research projects will be open knowledge all under a Creative Commons license (open code published on our GitHub repository and open data published on our Internet Data Portal (IDP)). 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.

Project status

The following sections give the status of research projects. Full details of each project are given in the Project Summaries section of this document.

In progress

Project	Progress this Quarter
ANZIC Classification of the register	Manual classification of more diverse set of training data. Close to completion.
Domain Retention Prediction	Reporting of results at registrar conference and in blog posts. Model has been successfully used by CENTR to model .be registry growth.
Domain Popularity Algorithm	Investigated effects of different DNS parameters on popularity calculation. Full traffic data from a .nz overseas DNS provider is being sought to check bias.
.nz HTTPS scan	Regular process running every second month.

On hold

Project	Next Steps
NZ Topology Map	The projects “NZ IP Topology Map” and “NZ BGP Topolgoy Map” will be combined in one, showing a single view from different data sources.
DGA Detection Algorithm	Proof of concept needs to be tested at a larger scale.
Register Word Decomposition	Requires a valid Te Reo Māori corpus to increase accuracy.
Full Web Scan of .nz	A proof of concept for the deep scan is available.
Zone Scan v2	Developer time required to replace the zone scan engine.

Complete

Project	Final Outputs
Registrant Classification	Working model complete and ready to be deployed into production.
DNS Statistics Publication Using IDP	Basic DNS stats uploaded into IDP. Blog post written interpreting results of scan.

NZRS an InternetNZ company

Project Summaries

Title	NZ Topology Map		**Status**	On Hold
Description	Mapping the internal structure of the Internet in New Zealand, using a combination of active and passive data collections. Passive data comes from BGP feeds from RouteViews, RIPE and Internet Exchanges. Active data collection uses RIPE Atlas probes in the country.			
Potential uses	There are a few outstanding questions about the structure of the NZ Internet whose answers can drive useful policy debate. For example, are there 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. Presented the New Zealand Internet view at NZNOG 2017.			
Collaborators	No active collaborators at this time.			
Progress	Needs work to run a regular collection. Make the raw data available via IDP.			

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 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. A combination of domain classifiers based on this work is being prepared for the Domain Analytics project.		
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. Preliminary results presented at the "Domain Usage Session" at CENTR Jamboree 2017.		
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	As this work is being integrated with Domain Analytics, all efforts have been concentrated on obtaining higher accuracy and training more models. There are now 7 models derived from this work in Domain Analytics with the following objectives: • to distinguish if a domain name is parked or not • to determine if it's For Profit or Non-for Profit • to determine the domain name's economic activity based on ANZSIC at three different levels, or determine the domain name's Non-for Profit activity at two levels. Additionally, work has been done to improve data collection from web scans - solving several issues that made them unreliable. In the short term, efforts will be focused on making the models run on a distributed platform. The original dataset acquired was not accurate enough. A larger and more diverse set of domains was therefore manually classified. New models have been trained using this dataset and work is close to completion.		

Title	Domain Retention Prediction	**Status**	On Hold
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. Results were presented at Registrar Conference 2017		
Collaborators	Some of the insights obtained in this work has been shared and discussed with staff at .CA. People from .IE (Ireland), Netherlands (.NL), and Austria (.AT) are following up this work closely.		
Progress	The combined forecast model to estimate the register size has been refreshed with new data. A new tool made available by Facebook has been incorporated that speeds up the computation and some of the decisions around outliers. The new model is accurately tracking 2017. The individual domain retention model is currently on hold, waiting for data from the register to be available. The model has been shared with CENTR and successfully used to model the .be registry.		

Title	Registrant Classification	**Status**	Complete
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	The results were presented at the Registrar Conference 2017 and it was voted as the best presentation.		
Collaborators	None.		
Progress	The Deep Learning solution requires more training data and efforts are focused on hand classifying 30,000 unique registrant names. A working model is available to be deployed to production.		

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 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. This work has been integrated into Domain Analytics.		
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	An experiment was designed to determine how different DNS parameters and software implementation affect DNS traffic volume, to account for that effect in the calculations. Full access to traffic from one of the .nz DNS offshore providers is being collated to provide us with data to check for bias.		

Title	DGA detection algorithm	**Status**	On Hold
Description	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.		
Potential uses	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.		
Planned outputs	*Form*	*Done*	*Details*
	Report	☐	
	Open code	☐	Will be published on GitHub.
Presented	The concept was presented at the New Zealand Internet Research Forum 2015.		
Collaborators	Details have been exchanged with SIDN Labs as they are working in similar ideas.		
Progress	The proof of concept needs to be tested at a larger scale, possibly using a different language.		

NZRS an InternetNZ company

Title	Register word decomposition	Status	On Hold
Description	Decompose every domain in the register into their word components (aucklandaccountants.org.nz into “Auckland accountants”).		
Potential uses	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.		
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	None.		
Progress	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.		

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.		

NZRS an InternetNZ company

Title	Zone Scan V2	Status	On Hold
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	Missing tests have been added and require testing.		

Title	DNS statistics publication using IDP	Status	Complete
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	☒	Blog post showing long term trends in the data.
	Open data	☒	Data available in IDP.
	Open code	☐	Will be published on GitHub.
Presented	Some high level statistics were presented at the Registrar Conference 2017. A private exchange with CZ.NIC comparing statistics for Q1 2017 at .cz and .nz.		
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	Basic DNS stats covering 2017 has been produced and uploaded to IDP. DNS stats upload is happening automatically. Additional statistics will be produced in the future. A blog has been posted analysing these statistics.

Title	.nz HTTPS scan		**Status**	In Progress
Description	Using our growing expertise on scanning the .nz namespace for data, we prepared a scan covering all active .nz domains and checked for HTTPS support and other related features.			
Potential Uses	There is increasing interest from the security community to understand how prevalent HTTPS support is in New Zealand. This data also gauges the presence of Certificate Authorities, adoption on new protocol features, and the operators' reaction to recent discovered vulnerabilities.			
Planned outputs	*Form*	*Done*	*Details*	
	Report	☐		
	Web site	☐		
	Open data	☒	Available in IDP https://idp.nz/Domain-Names/-nz-SSL-scan-results/cmxt-74aq	
	Open code	☐	Scanning code to be published on our Github account	
Presented	Summaries presented by Barry Brailey, Manager Security Policy for DNCL, at APRICOT 2017. Summary of results presented at the Registrar Conference 2017.			
Collaborators	None.			
Progress	Initial collection and data processing completed. Regular process running every second month.			

NZRS an InternetNZ company

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.
NZNOG	New Zealand Network Operators Group, a NZ-based organization gathering network operators, government and academy that organizes an annual meeting.
Hadoop	Big Data Platform
Deep Learning	Branch of Machine Learning using a set of algorithms that attempt to discover high level abstractions of the data.