

# Technical Research Report

## Introduction

This is April 2017 issue of the Technical Research Report, setting out the research projects being undertaken by NZRS and their state of progress. An updated version of this document is provided 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.

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 many 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	On Hold
<b>Description</b>	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.			
<b>Potential uses</b>	There are a few 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?			
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>	
	Web site	<input checked="" type="checkbox"/>	Website at <a href="http://ip.topology.net.nz">http://ip.topology.net.nz</a> updated with new version.	

	Open data	<input checked="" type="checkbox"/>	Resulting network representation made available via the project's website.
	Open code	<input checked="" type="checkbox"/>	Code available in NZRS GitHub account.
<b>Presented</b>	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.		
<b>Collaborators</b>	No active collaborators at this time.		
<b>Progress</b>	Needs work to run a regular collection. Make the raw data available via IDP.		

Title	NZ BGP Topology Map		Status	On Hold
<b>Description</b>	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.			
<b>Potential Uses</b>	Understand how the structure of the Internet in New Zealand changes with the pass of time, how different IXs gain/lose peers, etc.			
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>	
	Report	<input type="checkbox"/>		
	Web site	<input type="checkbox"/>	<a href="http://bgp.topology.net.nz">http://bgp.topology.net.nz</a> A new faster version will be made available soon.	
	Open data	<input checked="" type="checkbox"/>	Collected data made available via IDP	
	Open code	<input checked="" type="checkbox"/>	Code available in NZRS Github account	
<b>Presented</b>	Presented at First NZIRF and previous version at NZNOG 2014.			
<b>Collaborators</b>	AKL-IX, a relatively new Internet Exchange operator in New Zealand is willing to cooperate on feeding us their data for the map.			
<b>Progress</b>	Requires work to make it a regular collection.			

Title	ANZSIC classification of the register	Status	In progress
<b>Description</b>	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.		
<b>Potential uses</b>	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.		
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>
	Open data	<input type="checkbox"/>	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants
	Open code	<input type="checkbox"/>	Will publish code on GitHub
<b>Presented</b>	Concept presented at 2015 Registrar conference.		
<b>Collaborators</b>	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.		
<b>Progress</b>	<p>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:</p> <ul style="list-style-type: none"> <li>• to distinguish if a domain name is parked or not</li> <li>• to determine if it's For Profit or Non-for Profit</li> <li>• 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.</li> </ul> <p>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.</p>		

<b>Title</b>	Domain Retention Prediction		<b>Status</b>	In Progress
<b>Description</b>	Project to generate a probabilistic model that will tell us: <ul style="list-style-type: none"> <li>• Which elements of a registration are best predictors of their likelihood to be stay in the register</li> <li>• 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</li> </ul>			
<b>Potential uses</b>	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.			
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>	
	Report	<input checked="" type="checkbox"/>	A couple of blog posts are published in NZRS's blog.	
	Open code	<input type="checkbox"/>	Will publish code on GitHub.	

<b>Presented</b>	Concept presented at Registrar Conferences in 2014 and 2015. Results to be presented at Registrar Conference 2017
<b>Collaborators</b>	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.
<b>Progress</b>	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 tracking 2017 activity with good accuracy.

Title	Registrant Classification	Status	In Progress
<b>Description</b>	Machine Learning classifier to determine if a registrant is a person or an organization based on the registrant name.		
<b>Potential uses</b>	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.		
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>
	Open data	<input type="checkbox"/>	Will consider aggregated and anonymised data on IDP.
	Open code	<input type="checkbox"/>	Will be published on GitHub.
<b>Presented</b>	To be presented at Registrar Conference 2017.		
<b>Collaborators</b>	None.		
<b>Progress</b>	The Deep Learning solution requires more training data and efforts are focused on hand classifying 30,000 unique registrant names.		

Title	Domain Popularity Algorithm	Status	In Progress
<b>Description</b>	Algorithm using DNS data to determine if a domain name is more popular than others.		
<b>Potential uses</b>	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.		
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>
	Report	<input type="checkbox"/>	
	Web site	<input checked="" type="checkbox"/>	Some selected data sets are publicly visualized at <a href="http://domain-rank.nzrs.net.nz/popular.html">http://domain-rank.nzrs.net.nz/popular.html</a> and <a href="http://domain-rank.nzrs.net.nz/bank.html">http://domain-rank.nzrs.net.nz/bank.html</a>

	Open data	<input type="checkbox"/>	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants
	Open code	<input type="checkbox"/>	Will be published on GitHub.
<b>Presented</b>	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.		
<b>Collaborators</b>	Seeking collaboration within the CENTR group, as suggested by the CENTR R&D Chair.		
<b>Progress</b>	An experiment was designed to determine how different DNS parameters and software implementation affect DNS traffic volume, to account that effect into the calculations. Full access to traffic from one of the .nz DNS offshore providers will provide us with data to check for bias.		

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

Title	Register word decomposition		Status	On Hold
<b>Description</b>	Decompose every domain in the register into their word components (aucklandaccountants.org.nz into "Auckland accountants").			
<b>Potential uses</b>	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.			
	<i>Form</i>	<i>Done</i>	<i>Details</i>	

<b>Planned outputs</b>	Report	<input type="checkbox"/>	
	Open data	<input type="checkbox"/>	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants
	Open code	<input type="checkbox"/>	Will be published on GitHub.
<b>Presented</b>	None.		
<b>Collaborators</b>	None.		
<b>Progress</b>	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.		

<b>Title</b>	Full web scan of .nz		<b>Status</b>	On Hold
<b>Description</b>	Capture web content published under .nz domains to feed the ANZSIC classification project. Investigate tools to do a deeper gathering of content.			
<b>Potential uses</b>	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.			
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>	
	Report	<input type="checkbox"/>		
	Open data	<input type="checkbox"/>	Will be published openly on IDP but in aggregated form to preserve the privacy expectations of registrars and registrants	
	Open code	<input type="checkbox"/>	Will be published on GitHub.	
<b>Presented</b>	None			
<b>Collaborators</b>	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.			
<b>Progress</b>	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.			

<b>Title</b>	Zone Scan V2		<b>Status</b>	On Hold
<b>Description</b>	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.			
<b>Potential uses</b>	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.		
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>
	Open data	<input type="checkbox"/>	Will publish aggregated and anonymised data on IDP.
	Open code	<input type="checkbox"/>	Will be published on GitHub.
<b>Presented</b>	None		
<b>Collaborators</b>	IIS, the .SE register are collaborators as authors of the engine currently in use, and developers of the replacement.		
<b>Progress</b>	A test environment has been setup to run the unmodified version of the tool. This will explore the differences between the old and the new tool and some tests need to be added. This work has been allocated to the new NZRS developer for implementation.		

<b>Title</b>	DNS statistics publication using IDP		<b>Status</b>	In Progress
<b>Description</b>	Make data from the DNS traffic for .nz available using the Internet Data Portal			
<b>Potential uses</b>	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.			
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>	
	Report	<input type="checkbox"/>		
	Open data	<input type="checkbox"/>	Will publish aggregated and anonymised data on IDP.	
	Open code	<input type="checkbox"/>	Will be published on GitHub.	
<b>Presented</b>	A glimpse of the potential of this work will be presented at Registrar Conference 2017.			
<b>Collaborators</b>	SIDN is publishing some interesting counters from their DNS data, using a platform powered by Hadoop, inspired by the work we did with Hadoop			
<b>Progress</b>	Basic DNS stats covering 2017 has been produced and uploaded to IDP. A blog report will be published shortly to explore the data and document the work.			

<b>Title</b>	Digital Journey publication using IDP	<b>Status</b>	Finished
<b>Description</b>	Make data collected from the Digital Journey website about businesses self-assessment of their use of digital technologies available using the Internet Data Portal		

<b>Potential Uses</b>	Data collection started in 2014, and could provide a consistent view on how businesses have evolved their preparedness around digital technologies.		
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>
	Report	<input type="checkbox"/>	
	Web site	<input type="checkbox"/>	
	Open data	<input type="checkbox"/>	Available in IDP <a href="https://idp.nz/Users-and-Use/Digital-Journey/sp2s-ukz9">https://idp.nz/Users-and-Use/Digital-Journey/sp2s-ukz9</a>
	Open code	<input type="checkbox"/>	
<b>Presented</b>	None.		
<b>Collaborators</b>	MBIE as drivers of the initiative, Firebrand as developers and maintainers of the website.		
<b>Progress</b>	Initial upload of data completed with data from March 2014 to July 2016. Monthly updates scheduled.		

<b>Title</b>	.nz HTTPS scan		<b>Status</b>	In Progress
<b>Description</b>	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.			
<b>Potential Uses</b>	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.			
<b>Planned outputs</b>	<i>Form</i>	<i>Done</i>	<i>Details</i>	
	Report	<input type="checkbox"/>		
	Web site	<input type="checkbox"/>		
	Open data	<input checked="" type="checkbox"/>	Available in IDP <a href="https://idp.nz/Domain-Names/-nz-SSL-scan-results/cmxt-74aq">https://idp.nz/Domain-Names/-nz-SSL-scan-results/cmxt-74aq</a>	
	Open code	<input type="checkbox"/>	Scanning code to be published on our Github account	
<b>Presented</b>	Summaries presented by Barry Brailey, Manager Security Policy for DNCL, at APRICOT 2017.			
<b>Collaborators</b>	None.			
<b>Progress</b>	Initial collection and data processing completed. Requires work to run on a regular basis.			

## 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 <a href="https://idp.nz">https://idp.nz</a>
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.

INTENTIONAL BLANK PAGE