InternetNZ

New Report: Technical Research

In 2014 NZRS established a technical research function after Council identified this as an under-resourced area of InternetNZ's overall scope. To provide visibility of the work of this team, its goals and its outputs, NZRS has at Council's request prepared the attached February 2016 report on technical research.

As the report explains, the NZRS technical research team aims to expand the frontiers of our knowledge about the Internet within New Zealand and make that new knowledge openly available to all. Projects are drawn from the wide range of topics within this broad ambit.

This report includes:

- Descriptions of the current and proposed technical research projects as of February 2016
- The planned outputs of this research and the methods for sharing those;
- Collaborations and timings that may be possible with this research.

This is but the first of what we intend will be quarterly updates on the technical research programme.

Please contact me at Jordan@internetnz.net.nz should you wish to share your thoughts about any of these projects, or ideas of where we could go next. You are also very welcome to pass feedback on to Jay Daley, NZRS CEO, direct: jay@nzrs.net.nz and Sebastian Castro, Technical Research Manager, at sebastian@nzrs.net.nz .

Best,

Jordan Carter
Chief Executive

1 March 2016

NZRS an InternetNZ company

Technical Research Report

Introduction

This is the February 2016 Technical Research Report, setting out the research projects being undertaken by NZRS. This a new report, produced for the first time, provides better and more timely information on current projects and will be provided for 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?		
	Form	*Done*	*Details*

Planned output	Web site	☒	Existing website at http://ip.topology.net.nz.
	Open data	☒	Resulting network representation made available via the project's website.
	Open code	☒	Three code repositories on Github.
Presented	Proof of Concept presented at First NZIRF. Working version presented at Second NZIRF. Submission planned to RIPE 72.		
Collaborators	Active collaboration with RIPE Labs. RIPE has a tool that operates under the same principle of active measurements using the same measurement platform. There is interest in merging both projects and turn them into a framework to be available for other researchers and the community in general.		
Progress	All the data generating code is completed, tested and made public. The visualization code is tested and public, but still NZ-centric. Certain interesting analytics are available, like path anomaly detection (traffic destined to the country leaving the country). A new version of the frontend is ready for deploying that provides a better user experience.		

Title	ANZSIC classification of the register	**Status**	On hold
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 algorithm and be able to classify every domain in the register. This allow us to augment our understanding of the register		
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.		

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*
	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 (Registry for Canada)		
Progress	An approach using Machine Learning algorithms has been tested. Waiting for a suitable data dump from the SRS Operations Team to test if methodology produces meaningful results.		

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 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	None.		
Collaborators	None.		
Progress	A Proof of Concept is available as a web service, achieving a 95% percent accuracy. Requires discussion with the Operations Team to run the algorithm in a continuous basis.		

Title	Domain Popularity Algorithm	Status	On Hold

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 publish code on GitHub
Presented	Presented as Proof of Concept at DNS-OARC 22 in Amsterdam		
Collaborators	None.		
Progress	Algorithm needs testing for correctness and 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 publish code on GitHub
Presented	The concept was presented at the New Zealand Internet Research Forum 2015.		
Collaborators	None.		
Progress	The proof of concept needs to be tested at a larger scale, possibly using a different language.		

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 a sentiment analyser, identifying prevalence of geographic terms (and thereby understanding potential for a new geographic TLD) 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 publish code on GitHub
Presented	None.		
Collaborators	None.		
Progress	There is a Proof of Concept ready that needs to be tested for accuracy. One that's completed, identify how to put it into production.		

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 a other 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 publish code 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 needed to handle the cases where sites require Javascript to render content. A Proof of Concept for the deep scan is available using Hadoop, Heritrix and HBase.		

<table>
<tr><th>Title</th><th colspan="2">Zone Scan V2</th><th>Status</th><th>Not started</th></tr>
<tr><td><b>Description</b></td><td colspan="4">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.</td></tr>
<tr><td><b>Potential uses</b></td><td colspan="4">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.</td></tr>
<tr><td rowspan="3"><b>Planned outputs</b></td><td><i>Form</i></td><td><i>Done</i></td><td colspan="2"><i>Details</i></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 publish code on GitHub</td></tr>
<tr><td><b>Presented</b></td><td colspan="4">None.</td></tr>
<tr><td><b>Collaborators</b></td><td colspan="4">IIS, the .SE register are collaborators as authors of the engine currently in use, and developers of the replacement.</td></tr>
<tr><td><b>Progress</b></td><td colspan="4">Not started</td></tr>
</table>

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.