IAG T1 comments – Distribution of TMCH Data
January 30, 2012
Tom Barrett - EnCirca

The main topic of discussion for the IAG last week was concerned with distribution of TMCH data.

After comparing estimates of the maximum volumes for the TMCH to the query volume, uptime and responsiveness of existing registries, it seems very clear that the TMCH data could be kept solely in a central database utilizing a technical architecture similar in scale to some of the existing smaller gTLD registries, such as .MOBI and .TEL. This level of architecture for the TMCH could ensure the uptime and responsiveness required by registries and registrars while satisfying the trademark owner concerns for data security.

We agree with and confirm the following:
* The TMCH data is proprietary to trademark owners and is not currently in the public domain
* Prevention of data mining should be a priority for the TMCH
* Distributing the TMCH data makes data mining impossible to prevent
* Distribution of the TMCH data is not even necessary given the expected scale of the TMCH

Is TMCH Data Unique?
Trademark owners have said that the trademarks they have chosen to protect with the TMCH reflects a proprietary business strategy that they would want to keep away from their competitors and possible bad-actors. If the TMCH distributes data to registries and registrars for the processing of sunrise and trademark claims, then the potential of data abuse is magnified. Especially since competitors may well be registries themselves.

The people best positioned to determine if this information is proprietary or, at minimum, useful business information that they do not want distributed to competitors or potential bad actors, are the trademark owners. Since they believe TMCH submission data is proprietary and since the TMCH is a vehicle to protect trademark rights, deference should be given to their concerns.

Skeptics have challenged whether the TMCH data is truly proprietary or whether adequate controls could even be implemented to protect against data mining, given the need for registries and registrars to have some involvement during the sunrise and claims processes.

It has been pointed out that many government trademark databases are freely accessible. So, any determined competitor could regularly check these databases and view the activity of their competitors. There are also third-party trademark providers, such as Thomson-Compumark that, while not cheap, provide powerful searching and monitoring tools that can be used to determine the filing strategies of their competitors, especially across multiple jurisdictions.

Knowing a company's registered trademarks is valuable information to a competitor, and others, such as investors and domain speculators. But knowing which of these registered trademarks those companies have selected to protect via the TMCH is even more valuable information that cannot be determined from mere public trademark filings.

It is clear that a brand owner's TMCH strategy is proprietary and not available in the public domain.

Mining the TMCH Data

When confidential data must be kept secure, there are best practices for safeguarding and preventing data abuse. The primary principle of securing access to confidential data is to avoid making copies of the data. However if the data is freely distributed, data security techniques will essentially be useless.

We should recognize that the TMCH data, even if centralized, will be subject to data mining. There is little that can be done to stop some determined user from: a) searching public trademark databases to detect new filings of their competitors and then; b) methodically entering every string into their favorite registrar to see if a claims notice is generated? And if they do this, would we even know about it?

Although there is nothing to prevent a user from bulk searching at a registrar for the availability of domain names, we can prevent and detect data mining by incorporating some features related to how the claims notices are distributed, displayed and logged.

The most obvious steps we can take to protect the TMCH data include:

1. **Don't distribute the data.** If the claims data is distributed to registries and registrars, there is little chance that data mining can be prevented or detected. Imposing contractual penalties for abuse or misuse of the data would be ineffective, since once the data is leaked; there would be no way to determine who leaked it.

   The issue of a "single-point-of-failure" has been raised if the TMCH data is centrally located. But there are many examples of systems that are relied upon by tens of thousands of users and service providers that never see any downtime. So, how big of a TMCH would be required to process transactions with zero downtime and minimum lag time? I've have done such an analysis and presented the results below.

2. **Don't distribute owner-identifying data.** The claims process involves two basic steps: a) checking to see if an unregistered domain matches a record in the TMCH and b) retrieving the claims notice when there is a match. Step "a" does not include any owner-identifying information and could be distributed with a lower risk of data mining. But this may not be necessary or desirable given the scarce resources inherent in future mini-registries. Step "b" contains the more sensitive owner information and should not be distributed. For more, see my IAG comments on 12/12/2011.

3. **Log the display of claims notices.** The current spec includes the requirement to record when a registrant decides to proceed with a registration after being informed about existing claims. The trademark owner also receives an email when this event occurs.

   But the TMCH should also record the display of the claims notice itself, even if no registration follows. Obviously, since there was no registration, there would be no logging of who the potential registrant was. But we could log the string checked, the registrar used, the registry used and perhaps even the IP address used. All this data would be important historical data to help the community determine the effectiveness of the claims RPM.

   This data could also be used to detect possible data mining. It could help catch clever bad-actors who utilize multiple registrars and registrars to gather their data. Logging and collecting such data would be difficult if the data and responsibility is distributed to registries. It could also impose a new burden on new registries and registrars that they may not be equipped to deal with. There might even be an ancillary service here for trademark owners who wished to receive a report showing how

often their claims notices were displayed to potential registrants (who better to determine if their data is being mined?)

Issues with Distributing Data

The idea behind data distribution is that the TMCH transfers data and corresponding responsibility for processing sunrise registrations and processing trademark claims to registries and registrars. Each registry would then be held responsible for securing the data against abuse and ensuring the standards and procedures for sunrise and trademark claims are followed.

The advantage of this approach is that the TMCH is removed from any involvement for run-time sunrise and trademark claims, lowering the need for a high-availability technical infrastructure for the TMCH.

The disadvantage is that responsibility for sunrise and trademark claims now falls on thousands of registries. In addition to data security, data distribution raises other complexities for the TMCH. Such as, how do you ensure thousands of registries actually implement and perform the sunrise and claims tasks consistently, if at all? Would registrars be faced with different implementations of claims to deal with, potentially confusing end-users?

The reality is that, in the future, there will be hundreds of small-sized registries on the scale of .TRAVEL or .COOP. They will have limited revenues, staff and technical resources. Adding to their responsibilities vis-à-vis the TMCH, may not be the prudent thing to do.

ICANN compliance is also simplified by having the TMCH accountable for the RPM's, as opposed to thousands of registries and registrars.

Distribution of TMCH Data is Not Even Necessary

Since there appears to be little guarantee for data security with a distributed approach, we more closely examined the feasibility of the centralized data approach.

With a centralized approach, the TMCH would have these types of activity:
1) Authorizations for sunrise, likely via EPP
2) Queries of registrable strings to see if there is a matching claim, similar to Whois
3) Queries to retrieve claims notice when there is a matching claim, similar to Whois

Our analysis of a maximum volume scenario for the TMCH shows the following:

1. The maximum expected TMCH volume to process sunrise transactions is about double of today's .TRAVEL registry.

2. The maximum expected TMCH volume to process claims checks is on the same scale as today's .MOBI registry.

3. The maximum expected TMCH volume to display claims notices is on the same scale as today's .NAME and .TEL registries.

All of these existing registries typically manage to provide Whois access on a nearly 100% uptime on a 7x24x365 basis with a response time of less than 1500 milliseconds. If the selected TMCH vendor can demonstrate they can exceed these standards, then the need to distribute TMCH data is moot.

TMCH Capacity Analysis

This maximum "worst-case" scenario for the TMCH makes these assumptions:

1. There are 1,000 new gTLD's per year, matching ICANN's limit of adding 1,000 zones/year.

2. 167 new gTLD's are engaged in sunrise or trademark claims during any given 60-day period.

3. The average for sunrise registrations/TLD is 10,000 with a ratio of 2:1 for queries vs. registrations.

4. The average number of registrations during the 60-day claims period is 50,000/TLD with a ratio of 5:1 for TMCH queries vs. matches.

Details of the analysis are shown below:

1. The maximum expected TMCH volume to process sunrise transactions is 55,556/day, about double the daily Whois volume of 27,244 for .TRAVEL.

2. The maximum expected TMCH volume to process claims checks is 694,444/day, slightly higher than the 511,493 daily Whois volume for .MOBI.

3. The maximum expected TMCH volume to display claims notices is 173,611/day, which falls between the daily Whois volume of 157,331 for .NAME and 179,886 for .TEL.

4. For comparison purposes, I've also included the daily Whois volume for .BIZ and .COM below.

Model for TMCH Capacity Sizing

Assumptions	TMCH-normal	TMCH-Max	Comparisons (from ICANN.org for Sept, 2011)					
			.COM whois	.BIZ whois	.MOBI whois	.TRAVEL whois	.NAME whois	.TEL whois
Trademark Claims								
number of tlds/year	500	1000						
ave # of tlds/month	42	83						
tlds in claims period	83	167						
ave registrations/TLD/60 days	50,000	50,000						
queries/registration	5	5						
claims tmch queries/60 days	20,833,333	41,666,667						
claims-related queries/day	347,222	**694,444**	146,400,000	**1,762,452**	**511,493**			
% generating claims notices	25.00%	25.00%						
claims notices generated/day	86,806	**173,611**					**157,331**	**179,886**
claims-related queries/day	434,028	868,056						
registered names in TLD			*99,850,728*	*2,187,342*	*1,066,877*	*26,796*	*227,823*	*269,679*
Sunrise								
number of tlds/year	500	1000						
ave # of tlds/month	42	83						
tlds in sunrise period	83	167						
ave registrations/TLD/60 days	10,000	10,000						
queries/registration	2	2						
sunrise tmch queries/60 days	1,666,667	3,333,333						
sunrise tmch queries/30 days	833,333	1,666,667						
sunrise-related queries/day	27,778	**55,556**				**27,244**		