A privacy framework for 'DNS big data' applications*

Cristian Hesselman, Jelte Jansen, Maarten Wullink, Karin Vink and Maarten Simon
SIDN**
Contact person: Jelte.Jansen@sidn.nl

In this paper, we discuss our privacy framework for applications that further increase the security and stability of the .nl domain by using 'DNS big data' for purposes such as automatically detecting botnets in DNS traffic. Our framework is unique (1) because it introduces privacy management to the use of DNS data and (2) because, to that end, it integrates legal, organisational and technical aspects of privacy management. The framework will be incorporated by design into the ENTRADA platform – the technical system for DNS big data applications that we are developing at SIDN Labs.

I. INTRODUCTION

As the operator of the .nl domain – the Netherlands' country-code domain on the internet – SIDN is constantly looking for new ways of further enhancing the domain's security and stability. Hence, we have for example introduced DNSSEC on a large scale for .nl domain names [23], we contribute to the AbuseHUB [19] and we undertake research and development in this field through our R&D team, SIDN Labs [18].

We believe that the security and stability of the .nl domain can also be enhanced by retaining the DNS data that we process every day and performing automated data analyses to detect threats and irregularities early. The DNS (Domain Name System) [20][21] is the system that translates domain names into IP addresses. For instance, www.example.nl is translated into the numeric address of the web server that hosts the associated website. Translation is necessary because internet-connected computers use numeric IP addresses to send each other information, whereas people usually find it easier to work with domain names that they can recognise and remember.

We envisage DNS data applications such as the automated detection of botnets that use the DNS (e.g. Feederbot [1] or Cutwail [2]), automated data exchange with the AbuseHUB as a basis for working with ISPs to disable botnets, auto-configuration applications that independently reconfigure name servers in the event of DNS traffic abnormalities (e.g. as a result of a DNS reflection attack [3][4][5] or a sudden traffic spike), security performance analysis of top-level domains and .nl name server distribution analysis (cf [6]), sophisticated dashboards for DNS operators and open data applications.

Such applications are possible because of the emergence of new technologies, such as Hadoop [7], which can handle 'big data', a paradigm based on the analysis and enrichment of very large data flows (i.e. flows of petabytes of data) involving complex relationships and requiring rapid processing [8]. Big data has been described as data characterised by the three Vs: volume, velocity and variability. Traditional data processing methods, such as relational databases, are inadequate for big data, particularly in terms of processing speed and analytical capacity [8].

One challenge for 'DNS big data' applications is that some of the DNS traffic that the .nl name servers process consists of personal data, in particular IP addresses and domain names for which users want the corresponding IP addresses. We have therefore developed a privacy framework so that we can define a privacy policy for each application. The goals of the framework are (1) to enable us to systematically weigh the contribution of a particular DNS big data application to the stability and security of the .nl domain against the associated impact on the privacy of .nl users and (2) to configure the technical system in a way that enforces the implemented privacy policies. The framework is unique because it introduces privacy management to the use of DNS data and because, to that end, it integrates legal, organisational and technical aspects of privacy management. We will incorporate the

This paper is a translation of "Een privacyraamwerk voor 'DNS big data'-toepassingen", the original Dutch version of the paper. It is available online at www.sidnlabs.nl.

***SIDN (www.sidn.nl) manages the internet extension of the Netherlands, .nl. As the Dutch national domain name registry, we enable internet users to safely use and register .nl-domain names anytime and anywhere. To that end, we operate the .nl zone of the Domain Name System (DNS). We handle over a billion DNS queries every day for more than 5,5 million registered .nl domain names. Over 1.8 million of those are secured with "secure DNS" (DNSSEC), making .nl the largest secured internet extension in the world. SIDN Labs (www.sidnlabs.nl) is SIDN's research and development team, which for instance works on new internet technologies and systems to further enhance the stability and security of the DNS.*

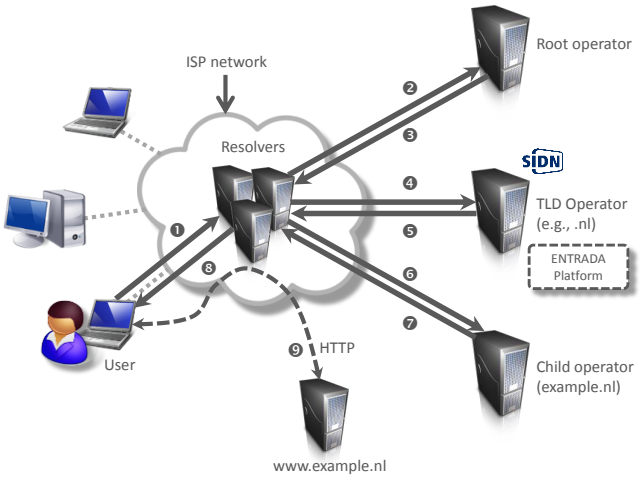

Figure 1. DNS resolving

framework 'by design' into the ENTRADA platform (ENhanced Top-level domain Resilience through Advanced Data Analysis), the technical system for DNS big data applications that we are developing at SIDN Labs.

As the operator of the .nl domain, we believe a sound privacy framework to be extremely important, because it is crucial in relation to confidence – both confidence in our national domain and confidence in SIDN as the operator of that domain. We also believe that we have a responsibility to be proactive in this field and to act transparently, because we provide a public service that is vital to the Dutch economy and Dutch society.

In the remainder of this paper, we discuss the ENTRADA privacy framework. We begin with a more detailed explanation of how the DNS works (Section II). Thereafter, we consider how and under what circumstances the IP addresses and domain names in the DNS traffic may constitute personal data (Section III). That is followed by the privacy framework itself (Section IV) and its realisation (Section 0). We end with a brief discussion of similar work (Section VI) and our conclusions and recommendations regarding further research (Section VII).

II. DNS RESOLVING

As the operator of the .nl domain, SIDN processes the messages that are exchanged when an internet user requires the IP address linked to a domain name, e.g. the IP address 192.0.2.189, linked to www.example.nl. Domain names need to be translated because internet-connected computers use numeric IP addresses to send each other information, whereas people usually find it easier to work with domain names that they can recognise and remember. The translation of domain names into IP address is known as 'resolving' and takes place via the Domain Name System (DNS) [20][21], a global server infrastructure, within which SIDN operates the .nl part.

Figure 1 shows how DNS resolving typically works. The process starts with a user typing a URL into his/her browser – let's say http://www.example.nl/ – or clicking on a link to that address. The part between the slashes (www.example.nl) is the domain name of the address and refers to the server that hosts the relevant website. In order to translate the domain name into the server's IP address, the user's machine sends a DNS query to a so-called 'resolver' (step 1 in Figure 1). A resolver is another machine, usually operated by the ISP through whom the user acquires internet access. In response to the browser's query, the resolver looks up the domain name in the global DNS, starting with a fixed group of 'root servers' (step 2). In the case of www.example.nl, the root servers refer the resolver to the name servers for .nl (step 3). The resolver accordingly contacts a .nl name server (step 4), which duly refers the resolver to the name servers for example.nl (step 5). The resolver then sends a DNS query to the name server for example.nl (step 6), which knows the IP address for www.example.nl and sends it to the resolver in a reply message (step 7). Finally, the resolver replies to the user's browser, giving the IP address for www.example.nl (step 8). The browser is then able to retrieve the web page at www.example.nl using HTTP (step 9).

In order to maximise the scalability of the DNS, resolvers 'cache' data. Caching involves storing a DNS reply for a certain period of time, so that, if another client asks for the IP address of the same domain name, the resolver can immediately reply from its own cache, without having to contact the name servers in the DNS again. In other words, the resolver skips steps 2 to 7. The length of time that a resolver keeps a reply depends on the 'time to live' (TTL) of the domain name. Within the TTL, an individual resolver will not normally ask the .nl name servers about a given domain name more than once. As the .nl registry, SIDN recommends a TTL of two hours. A resolver may use a

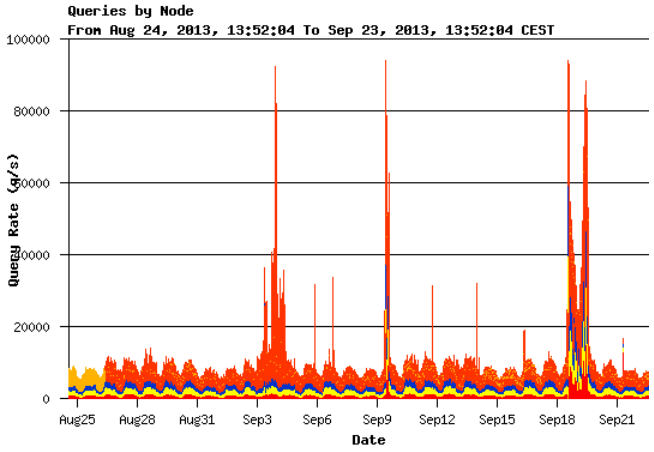

Figure 2. DNS queries handled by one of SIDN's unicast-name servers.

shorter TTL if its operator wishes, but not a longer one.

The average query rate for all .nl name servers combined (unicast and anycast) is about 15,000 queries per second, which equates to 39 billion queries (and responses) per month. If all the data was recorded (including the IP and Ethernet headers), under normal circumstances about 60 gigabytes of storage space would be required per day per name server. As an indication, Figure 2 shows the volume of traffic handled by one of SIDN's unicast name servers in the course of a month (August 2013).

III. PERSONAL DATA

Under certain circumstances, the IP address of a resolver and the domain name requested in a DNS query constitute personal data. In this section, we consider why some DNS data must be regarded as personal data in the sense of the Dutch Data Protection Act and how the relevant requirements of that Act are complied with (Section III.A). Using operational .nl data, we also estimate how often the IP addresses we process and domain names included in user queries constitute personal data (Sections III.B and III.C, respectively).

A. WBP analysis

The Dutch Data Protection Act (known by its Dutch initials, WBP) [15] defines personal data as 'any piece of information regarding an identified or identifiable natural person' and the processing of personal data as 'any action or sequence of actions involving personal data, including but not restricted to the collection, recording, sorting, […] deletion or destruction of such data' (Section 1, subsection a). On the basis of those definitions, we believe that the IP address of a resolver and the domain name requested in a DNS query (see Section II) can constitute personal data under certain circumstances.

Our interpretation is that, in such cases, other data contained in the DNS query processed by the .nl name servers should, by association, be regarded as personal data as well. The other information in question includes the query time stamp, protocol flags in DNS queries that provide information about the resolver, 'volatile' data such as transaction numbers for the query itself (query ID, source port) and the resolver's network details, such as the distance to the .nl server (network hops).

As the registry for .nl, we take the view that such data can be processed on the ENTRADA platform, because such processing satisfies the criteria for personal data processing set out in the WBP (purpose limitation, legitimate basis, conditions for processing special personal data and informing the subject).

Purpose limitation: personal data may be collected only for specific, explicitly defined and justified purposes (Article 7) and may not be processed in any way that is inconsistent with the purpose for which it was obtained (Article 9). Where the ENTRADA platform and its applications are concerned, the purpose of data processing is the prevention of fraud and abuse and the further enhancement of the stability of the .nl domain and the internet in general. We do not use the data for other purposes, such as commercial purposes.

Legitimate basis: the WBP also specifies that we may process data only with a legitimate basis (Section 8, subsection f). In the context of ENTRADA, the reason for processing is the pursuit of a legitimate interest. Combatting fraud and abuse and further enhancing the stability of the internet serve not only SIDN's legitimate interest in (promoting) the security and reliability of the .nl domain, but also, in individual cases, the legitimate interests of the parties to whom we make data available. Examples of such interests include the interests of the owner of an infected computer, whom we inform (through his/her service provider) about the infection, and the interests of an organisation that has come under a DDOS attack, which we provide with information that will help it fend off that attack. The processing of personal data does not conform to Article 8, sub f, of the WBP if 'the interest or the fundamental rights and freedoms of the subject, in particular the right to privacy, takes precedence'. Consequently, an assessment must be made regarding each ENTRADA application.

Special personal data: in principle, the WBP prohibits the processing of special personal data, such as a person's religion or philosophical belief (Section 16). However, Article 23, sub 2, makes an exception to allow the processing of such data for scientific research or statistical analysis, provided that the research or analysis serves a general interest, that the processing is necessary for the research or analysis in question, that seeking explicit consent is impossible or would involve disproportionate effort and that the research or analysis is organised so as to ensure that the subject's privacy is not disproportionately compromised. In very exceptional circumstances, the ENTRADA platform may process special personal data (Section III.C), but we never link the processed information to particular individuals. The processing does not interfere with the subject's privacy and therefore conforms to Article 23, sub 2.

Informing the subject: given the way that the DNS works (see Section II), we consider it impracticable for users to be informed via the service itself about the use of their data. DNS resolving takes place 'invisibly' within the internet infrastructure, meaning that we have no opportunity to seek subjects' consent interactively, as one may do before a user opens a website, for example. Our

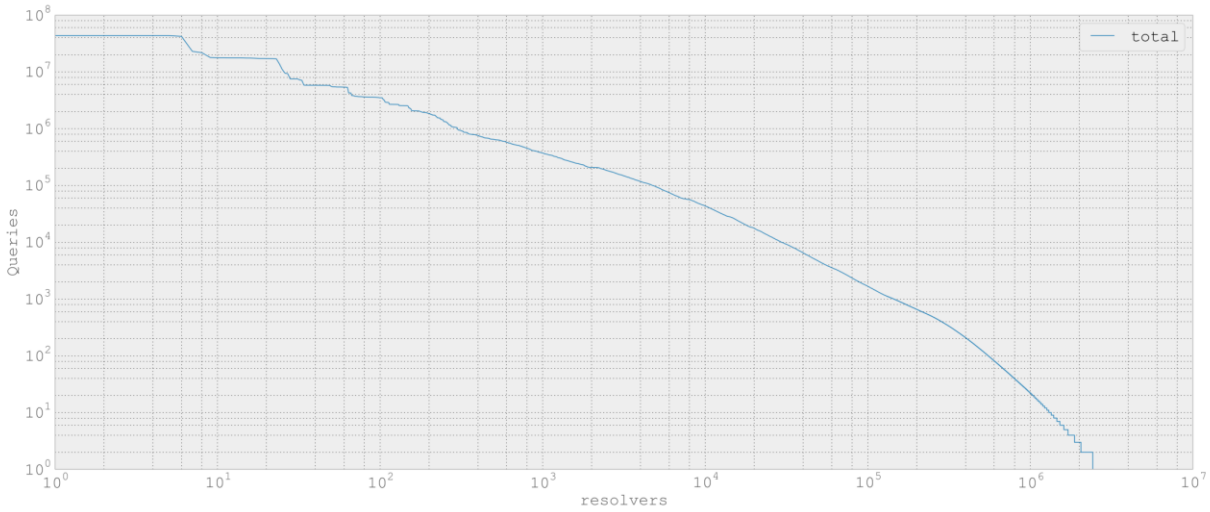

Figure 3. Average number of queries per day (June 2014)

interpretation of Article 34, sub 4, is that it is sufficient for us to record the origin of the data we collect.

B. IP addresses

The resolver IP addresses contained in the DNS queries that our name servers process constitute personal data if they can be traced to a natural person. In this context, we distinguish between two possible scenarios: (1) the resolver in question serves a small number of users (one or a handful) and (2) the resolver in question serves a large number of users. We consider it probable that the IP address of the resolver is personal data in the first scenario, but improbable that it is in the second scenario. In proceeding on that basis, we are following the WBP as currently worded and setting aside the question of whether all IP addresses should be regarded as personal data [9] until such time as the law states that they should.

In the first scenario, the resolver IP address handled by the name server is *probably* personal data when combined with the domain name that an individual user is looking up. Although the name server sees the resolver's IP address and not the end user's IP address, the small number of users means that the resolver is likely to consult the name servers comparatively often (cache misses), with the result that the name servers see a relatively large proportion of the domain names looked up via that resolver. That may be the case if, for example, a family has a local resolver at home.

The IP address of a resolver is also probably personal data if the IP address of the user and that of the resolver are the same. That situation can arise in a botnet, which uses its own resolver to control infected machines. An infected machine consults the botnet's resolver instead of the resolver that it would ordinarily use (e.g. the resolver assigned by the user's ISP). That is the case, for example, with the Feederbot botnet [1] and the Cutwail botnet [2]. The same situation also arises where the operating system allows the user to operate a local resolver. That is often the case with operating systems used by expert users, such as FreeBSD10 [11] and DNSSEC Trigger [12].

In the second scenario (a resolver that serves a large number of users) it is *improbable* that the IP address is personal data. Again, the name servers see only the resolver's IP address, but now in combination with a relatively small proportion of the domain names queried via that resolver. That is because the greater traffic volumes enable the resolver to answer more queries from its cache (without steps 2 to 7 in Figure 1), certainly where the most popular domain names are concerned. Consequently, only a small portion of an individual user's DNS queries typically reach the name servers. An individual user who looks up a large number of unique domain names not visited by other users served by the same resolver is potentially recognisable from name server data, but it is improbable that the user is identifiable from an IP address.

We have performed an initial analysis of resolvers consulting the .nl name servers, to establish the breakdown between resolvers that serve a small number of users and resolvers that serve a large number. Establishing that breakdown is important, because the number of resolvers that serve a small number of users is a design determinant for our privacy framework, influencing matters such as configuration of the ENTRADA platform's data filters (see Section IV). We note that we can only estimate the number of users served by a resolver, because it is merely the externally observable behaviour of a resolver that is apparent to a name server (interactions 4 and 5 in Figure 1).

As an indicator of the number of users served by a resolver, we have used the number of queries received

from the resolver in question. Other potential indicators will be investigated in the context of future research (see Section VII). Examples include the distribution of domain names looked up by end users, interval between successive queries, the port numbers used and the number of network hops between resolver and authoritative name server. In order to determine the number of users more accurately, it would be necessary to have information about the network between the resolver and the end user.

Figure 3 shows the average number of queries per day handled by one of our production name servers over the course of a month (June 2014). The x axis shows the number of resolvers (3,211,225 in total) and the y axis the average number of queries per resolver per day. Both axes have a logarithmic scale. The dataset that we used contains more than 3.74 billion DNS queries collected and analysed using the R&D version (prototype) of the ENTRADA platform.

Figure 3 clearly shows an uneven distribution of resolvers across the query count range. Only a small number of resolvers send a large number of queries (left-hand end of Figure 3) and therefore probably serve a large number of users. The top hundred resolvers send 28 per cent of all queries received by the name servers and the resolver in hundredth place sends an average of 117,000 queries per day. The same pattern is evident in Figure 4, where the number of resolvers is plotted against the number of queries that we receive. Each column in Figure 4 indicates the number of resolvers that send between $10^{N-1}+1$ and 10^{N} queries (where N=1…8). The column on the far left is the number of resolvers that send a single query.

With the aid of reverse DNS, we are able to see that resolvers that send large numbers of queries often belong to large ISPs or large companies. It is therefore improbable that the associated IP addresses are personal data. However, the highest-volume resolvers include those operated by 'domainers', who seek to ascertain what domain names there are in the .nl zone. Domainers often use small numbers of automated systems, meaning that the associated resolver IP addresses may be personal data.

The great majority of the queries reflected in Figure 3

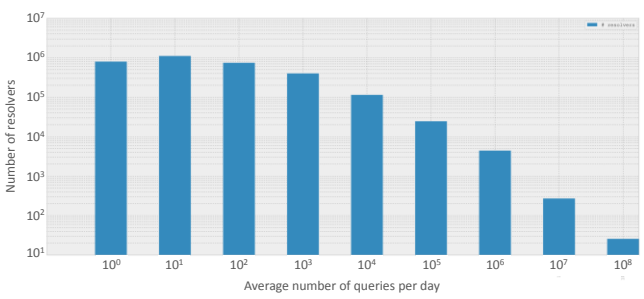

Figure 4. Number of resolvers per DNS query number interval.

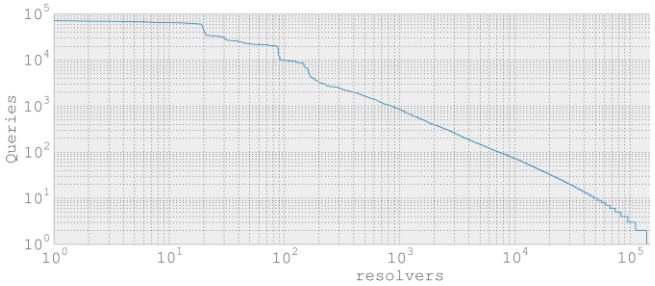

Figure 5. Number of DNS queries regarding domain names that contain IP addresses

originate from resolvers that each send a fairly small number of queries and therefore probably serve only a small number of users. So, for example, resolvers that send a hundred queries per day or fewer account for 83 per cent of the total. The resolvers in question probably each serve a small number of users or even run on end users' own machines. For the great majority of resolvers, therefore, the associated IP address is personal data.

C. The domain names that are looked up

The domain name that an end user looks up (see Figure 1) can itself constitute personal data. For instance, a domain name that incorporates the name of a natural person, e.g. www.firstname-surname.nl, is likely to be personal data – albeit relating to the registrant of the domain name, rather than to the person using the resolver. A platform such as ENTRADA could, by analysis of retained DNS query data, link other information to the domain name, which should then be considered to be personal data regarding the registrant. Additional information that constitutes personal data by association might include the IP addresses looking up the domain name, and the times and frequencies of the look-ups. It is questionable how significant such information might be, but that does not in our view alter the fact that the information conforms to the definition of personal data.

Without other information, a looked up domain name may also be personal data if it contains an IP address. ISPs sometimes incorporate IP addresses into domain names in order to identify the connection to the client. Examples that we have come across include:

- <IP address>.customer.<ISP name>.nl
- <IP address>-dsl.<ISP name>.nl
- <IP address>-mx.xdsl.<name ISP>.nl
- <IP address>-static-<ISP name>.dsl.ip.<ISP name>.nl

In such constructions, a hyphen is sometimes used to represent each dot between number groups in the IP address. Figure 5 shows how often such constructions occurred in queries received in June 2014 (y axis). There were about 100,000 resolvers that sent queries regarding

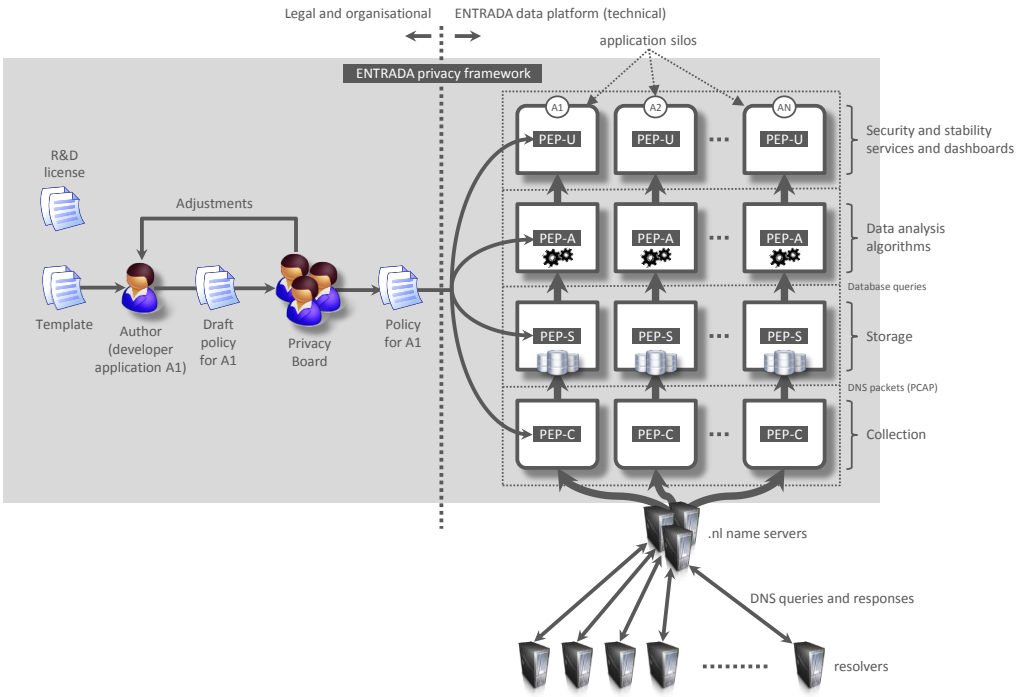

Figure 6. ENTRADA privacy framework (production).

domain names with embedded IP addresses (x axis), and the maximum number of queries sent was about 100,000 (y axis).

Finally, the domain name that is looked up is personal data if the IP address in the DNS query received from the resolver is personal data (see Section III.B). If, for example, the domain name that is looked up is linked to a gay website, that may say something about the person for whom the resolver is looking up the name. Hence, all the other data we obtain regarding the query – e.g. the time of the query – are personal data as well.

IV. ENTRADA PRIVACY FRAMEWORK

Figure 6 provides an overview of the ENTRADA privacy framework as we envisage it for future production services and applications. The central concept is that of a privacy policy, which defines what data the ENTRADA platform processes for a particular application, for what purpose and using which personal data filters. A filter is an operation performed on the personal data (e.g. pseudonymisation or aggregation) in order to adhere to the principles of proportionality and subsidiarity by avoiding excessive or unnecessary processing of personal data. The filters form an essential element in the privacy framework, because they ensure that the policies are verifiably enforced by technical means. Within the ENTRADA privacy framework, each application has its own privacy policy.

The ENTRADA privacy framework is based on requirements such as purpose limitation and verifiability (Section IV.A) and defines two organisational roles: policy author (Section IV.A) and privacy board (Section IV.C). Policy authors are usually the application's developers and define a privacy policy for a given ENTRADA application. The privacy board reviews the policies and approves realisation of the relevant application, with associated privacy policy. The technical componentry of the framework consists of policy enforcement points, which enforce the approved policies (Section IV.D), and so-called 'application silos', which ensure that data remains associated with the purpose for which it was collected (Section IV.E). Thus, the framework covers the legal aspects (the privacy policies), organisational aspects (the policy authors and privacy board) and technical aspects of privacy management (the policy enforcement points and application silos). We are also applying the privacy framework to the R&D version of the ENTRADA platform (Section IV.F), which we use for development of the privacy mechanisms for the production version, for example.

Some of the concepts that we use are inspired by the IETF's policy framework [22].

A. Requirements

Our starting point for the ENTRADA privacy framework is the analysis described in Section III. Our requirements for the framework are:

- *Purpose limitation*: the framework must link the use of personal data (resolver IP addresses or requested domain names) to a particular

ENTRADA application with a particular purpose and actively ensure that all use is linked to the defined purpose. The exclusive purpose of every ENTRADA application is to further enhance the stability and security of the .nl domain. In service of that purpose, the specific aim of the application may be, for example, the detection of botnet traffic in DNS data or the facilitation of research in the field by research centres. We explicitly prohibit the use of ENTRADA data for commercial purposes.

- *Verifiability*: the framework must consist of technical, organisational and legal concepts and structures, which enable SIDN to systematically determine which personal data a given ENTRADA application requires, for what purpose and what privacy measures may be necessary to ensure the legality of processing (e.g. anonymisation of IP addresses). Policy authors must, for example, be able to use the framework to define the personal data required by the relevant ENTRADA application.
- *Simplicity*: the framework and the documentation associated with its application must be as simple as possible, in order to ensure its utility for policy authors. Simplicity is also important for making the framework technically, legally and organisationally implementable, so that we may ultimately integrate it within our operational processes and systems. We currently do not consider extensive automation of the framework necessary, e.g. through the use of web forms, because we anticipate making only occasional use of the privacy framework, typically when developing or modifying ENTRADA applications. Furthermore, we believe that some human reflection is always desirable in the field of privacy and that over-automation should therefore be avoided.
- *Extensible*: the framework must be sufficiently generic to be usable if, for example, the European Data Protection Regulation [17] replaces the Data Protection Directive [16], on which the WBP is based. The privacy framework must also be capable of extension to other data flows that we process as operator of the .nl domain, with a view to enabling the early identification of threats and irregularities in those flows as well. For example, the framework should be applicable to Extensible Provisioning Protocol (EPP) traffic [13].

B. Policy authors

A policy author is someone who writes a privacy policy for an ENTRADA application. In terms of Figure 6, that implies that each application A1, A2 up to and including AN has a policy author who is responsible for management of the policy.

A privacy policy takes the form of a text document whose structure resembles that of the personal data processing declaration form provided by the Dutch Data Protection Authority (CBP) [10]. A privacy policy has the following elements (see Appendix B for a specimen policy):

- *Identifier*: identifies the policy.
- *Purpose*: description of the ENTRADA application that requires the personal data, the purpose to which the data will be put and the benefits for the stability and security of the .nl zone. That might be a monitoring application that serves to automatically detect botnets in DNS traffic as a basis for enhancing the security of the .nl domain.
- *Personal data*: the types of personal data to which the policy applies. Where DNS traffic is concerned, that may be 'IP address', 'requested domain name' or both (see Section III).
- *Filters*: a description of the filters used for this policy, the circumstances of their use and the types of personal data to which they are to be applied. Filtering options may include pseudonymisation, deletion of personal data and no filtering.
- *Retention*: the length of time for which we retain the personal data required by the application. At the end of the retention period, the ENTRADA platform deletes the personal data or retains it in anonymised form.
- *Access*: a specification of the people or systems that will have access to the data and of the conditions applying to such access. If a system is to have access, the policy must also describe the security arrangements and state how the system will obtain the data.
- *Type*: the policy definition must state whether the policy applies to an R&D activity or to an SIDN production service. Privacy policies for R&D activities (see Section IV.F) will usually involve few if any filters, with a view to facilitating the development and evaluation of the privacy mechanisms themselves, for example. Where production services are concerned, we know exactly which personal data are required and we can apply the privacy policies strictly.

- *Other security measures*: a description of any security measures not referred to under the other headings.

Policy authors make use of the ENTRADA policy template to write each new policy. The template is a text document, which defines the structure of a policy and provides information regarding each element. Use of the template ensures that all ENTRADA policies have a uniform structure and that their content is standardised as far as possible. That in turn simplifies the drafting of a policy, its evaluation by the privacy board (see Section IV.C) and its subsequent publication.

For each type of personal data (IP address or requested domain name) the ENTRADA policy template also defines the various filters available to policy authors for assignment to the application. The template describes the pros and cons of each filter, so that policy authors can make informed decisions about which filters to use (see Appendix A).

The ENTRADA policy template distinguishes four types of filter, one for each stage in the processing of personal data on the ENTRADA platform (see Figure 6): collection (of DNS traffic on the .nl name servers), storage (in a database), analysis (by algorithms and combination with other sources) and use (by ENTRADA services and applications). The filters available for IP addresses include: the omission of IP addresses when query data are collected (collection filter), the partial zeroing of IP addresses before the collected data is saved (storage filter), the aggregation of data so that source data more than, say, one day old are deleted (analysis filter) and the non-disclosure of IP addresses when data is accessed (usage filter).

We also apply privacy policies to applications that involve the sharing of data with third parties, such as the AbuseHUB [19], with a view to tackling botnets.

C. Privacy board

The privacy board is a body within SIDN, which is responsible for evaluation of the privacy policy for an ENTRADA application and which decides whether the purpose of the application warrants the means used (validation). To that end, the board weighs up the ENTRADA application's contribution to the stability and security of the .nl domain against the implications for the privacy of .nl users (the need to protect the fundamental rights and freedoms of the subject [15]). In Figure 6, the privacy board approves the privacy policy for application A1, after which the ENTRADA platform enforces the privacy policy technically.

The board is made up of a legal expert, a technical expert and a chairperson. The privacy board publishes approved privacy policies and the basis for their approval on SIDN's website. Publication maximises transparency and encourages the board to ensure that all its decisions are completely defensible.

The privacy board assesses privacy policies before any associated new service is taken into production. The board is additionally responsible for assessing revisions to existing privacy policies and for maintenance of the ENTRADA policy template, e.g. by updating the list of privacy filters.

D. Policy enforcement points

A policy enforcement point (PEP) is a software component of the ENTRADA platform, which serves to implement the privacy policy filters for a particular application.

We distinguish four types of PEP, one for each level of the ENTRADA platform (see Figure 6):

- PEP-C: implements privacy policies relating to the *collection* of DNS data, e.g. the deletion of IP addresses from the DNS data prior to storage. The PEP-C works directly on the queries and replies processed by the .nl name servers.
- PEP-S: implements privacy policies relating to the *storage* of DNS data, e.g. the aggregation of data after a given period of time. The PEP-O works on the databases in which we store the DNS data, such as Hadoop [7].
- PEP-A: implements privacy policies relating to the *analysis* of DNS data, e.g. the definition of analysis algorithms that produce only information that cannot be traced back to individuals.
- PEP-U: implements privacy policies relating to the *use* of DNS data by services and applications, e.g. the sharing of data with initiatives such as the AbuseHUB [19].

E. Application silos

In addition to the four layers of the ENTRADA platform (see Section IV.D), we also distinguish so-called 'application silos'. An application silo serves to ensure that the personal data in the ENTRADA platform always remains linked to the particular purpose for which it was collected (the application) and does not find its way into other silos linked to other purposes (applications).

In the ENTRADA platform, an application silo consists of the application, all the stored personal data required by the application and the associated privacy policies. The arrangement may be realised by, for example, configuring the ENTRADA platform so that each silo runs on its own (virtual) machine. That minimises the need for special technical measures within the platform, keeping the platform as simple as possible and enabling us to realise the separation of silos primarily on an organisational basis.

F. *Research and development*

The ENTRADA platform and the privacy framework concepts are highly innovative. They are dependent on thorough research and development (R&D) for purposes such as the development of algorithms to detect botnets in DNS traffic and the creation of new privacy filters.

We therefore make explicit distinction between production and R&D. In the production environment, we adhere strictly to the approach illustrated in Figure 6. For R&D, however, we employ a more flexible regime, because the purpose of R&D activities is to investigate how an ENTRADA application may be realised (e.g. using which mechanisms, algorithms and privacy policies), evaluated and afterwards possibly taken into production. The intention is not to make the application available to .nl users, except where pilot ENTRADA applications with which .nl users voluntarily cooperate are concerned.

Our R&D regime is more flexible than the production regime, in the following respects:

- We use application silos with 'porous walls'. That means that we share personal data between various applications within our lab environment, but exclusively for R&D purposes. We also share research data with our (academic) R&D partners, in which context we pseudonymise or anonymise the data where necessary prior to sharing. Such situations are regulated by processor agreements that we make with the parties concerned.
- We retain the DNS data (and the incorporated personal data) used for all ENTRADA applications for the same length of time, rather than for an application-specific period, as in production (see Section IV.B). Article 10, clause 2, of the WBP states that data may be retained for an extended period for historical, statistical or scientific purposes. No specific limit is given. We have chosen to use a retention period of 18 months, so that we have sufficient time to analyse the data for a whole year and to report our findings. Data more than 18 months old is anonymised or destroyed. The 18-month period is a 'sliding window', and the privacy board can incidentally authorise extension of the window for specific R&D purposes.

When an ENTRADA application is transferred from the R&D environment to production, it enters production use in an empty silo (i.e. without any DNS data). By means of that strategy, we ensure that any personal data used in the application's development cannot be used for another purpose (the provision of an ENTRADA-based production service instead of research and development).

V. Realisation

We are currently in the process of setting up the ENTRADA privacy framework. We will begin by informing the Data Protection Authority (the body charged with regulating personal data protection in the Netherlands, as provided for in Part 4 of the WBP) of our intentions. We are also working internally to establish the privacy board. The Board's responsibilities will be increased incrementally and assessment of the prototype ENTRADA platform will be used as an opportunity to test the arrangements.

Technically speaking, the ENTRADA platform has been designed on the basis of a 'plug-in' architecture. The filters required to enforce the defined privacy policies can therefore easily be added to the platform, thus achieving privacy by design. So, for example, a plug-in can be created to handle the anonymisation of IP addresses. The ENTRADA platform has been realised in the form of a Hadoop cluster.

The prototype of the platform currently operates only in a discrete laboratory environment. Given its experimental nature, we expect that it will be some time before we have an ENTRADA-like system for our operational services. The privacy framework has nevertheless been developed and organised in anticipation of that situation.

VI. Related work

Krishnan and Monrose [14] have written about the privacy implications of DNS prefetching. With DNS prefetching, a browser (e.g. Google Chrome) starts resolving the domain names that occur on a web page while the user is still entering a search term in the browser's address bar or while the page is loading. The advantage is that pages load more quickly, thus improving the 'browser experience' for the end user. Krishnan and Monrose's research shows that DNS prefetching highlights the privacy risks associated with the use of domain names. By resolving all the domain names on a page, the user's browser will quickly cause a great deal of extra context information to be added to the resolver's cache, potentially enabling a clear picture of the user's search query to be built up from the resolver data. The method is less practical if only a proportion of the queries reach the .nl name servers (due to caching by resolvers) or if the domain names used are made up of words that bear no relation to the search query. Krishnan and Monrose's research differs from ours insofar as those researchers concerned themselves only with the technical scope for privacy protection and disregarded the legal and organisational aspects.

Project Turris [24] is a service operated by CZNIC – the registry for the Czech country-code domain, .cz. Turris analyses (DNS) data traffic with a view to detecting and fending off attacks on the internet. To that end, it uses a

special router, which users install in their home networks. Because the service involves the analysis of end users' data traffic, CZNIC has developed a special privacy policy [25]. The Turris policy equates to a specific protection and aggregation policy in our framework. Turris differs from our ENTRADA privacy framework insofar as, by entering into a lease contract with CZNIC, end users give explicit consent for the use of their data in the Turris project. Another difference is that Turris involves the use of much more detailed data than that used for ENTRADA, since Turris gathers data on all the user's internet traffic, and that data may be traced back to individual users. Nevertheless, Turris is similar to ENTRADA in the way that data is recorded in a secure form and the way that aggregation and security measures are implemented before any data are shared with users.

Leenes has published an expert opinion [28] on the permissibility of the processing of botnet data by SURFnet, an organisation that connects the networks of universities and other academic institutions in the Netherlands. Leenes describes the data involved, the associated legal privacy considerations and the actions that may be performed. The dataset considered by Leenes does not consist of DNS data, but botnet data. As such, it is a more specific dataset, but one that includes a lot more personal data. The article focuses primarily on the legal aspects.

VII. Conclusions and Future Work

Applications that record and automatically analyse DNS data have the potential to further enhance the security and stability of the .nl country-code domain. We have developed a privacy framework for such applications, because some of the DNS data constitutes personal data (resolver IP addresses and the domain names looked up by users). Our framework is unique (1) because it ensures that 'DNS big data' applications provide privacy protection and (2) because, to that end, it integrates legal, organisational and technical aspects of privacy management. We believe that a thorough and transparent approach to privacy protection in this field is very important because the .nl domain forms a public infrastructure that is vital to the Dutch economy and Dutch society.

On the technical level, our future work will entail, for example, the development and evaluation of mechanisms for distinguishing more accurately between resolvers that serve a small number of users and those that serve a large number, e.g. by using combinations of indicators (see Section III.B) or by employing machine learning technology.

Where the legal aspects are concerned, we intend to investigate how we can share our (enriched) DNS data under a licence that incorporates (parts of) our privacy policy template. We will also examine the implications of transition from the European Data Protection Directive to the Data Protection Regulation. The existing WBP is the Dutch implementation of the Directive and the Data Protection Regulation is to be stricter in certain respects. The change will involve the replacement of a directive with a regulation with direct effect in member states. The proposals are still under discussion and it therefore remains unclear which provisions of the new legislation will differ from the current law.

Finally, we wish to additionally apply the privacy framework to other types of traffic that we handle in our role as country-code registry, such as EPP traffic.

Acknowledgement

We wish to thank Arnold Roosendaal (TNO), Simon Hania (TomTom and a member of SIDN's Supervisory Board) and the Privacy and Identity Lab research staff for their feedback on the draft version of this paper, on the basis of which various improvements have been made.

References

[1] C. Dietrich, C. Rossow, F. Freiling, H. Bos, M. van Steen, and N. Pohlmann, "On Botnets that use DNS for Command and Control", 7th European Conference on Computer Network Defense (EC2ND '11), Gothenburg, Sweden, September 2011, http://www.syssec-project.eu/m/page-media/3/dietrich-ec2nd11.pdf

[2] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vign, "The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns", 4th USENIX conference on Large-scale exploits and emergent threats (LEET'11), Boston, USA, March 2011, https://iseclab.org/papers/cutwail-LEET11.pdf

[3] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis", ISOC Network and Distributed System Security Symposium (NDSS 2011), San Diego, California, Feb 2011, http://www.syssec-project.eu/media/page-media/3/bilge-ndss11.pdf

[4] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon, "Detecting Malware Domains at the Upper DNS Hierarchy", 20th USENIX Security Symposium, San Francisco, California, Aug 2011, https://www.usenix.org/legacy/event/sec11/tech/full_papers/Antonakakis.pdf

[5] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dago, "From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware", 21st USENIX Security Symposium, Bellevue, WA, Aug 2012,

https://www.usenix.org/system/files/conference/usenixsecurity12/sec12-final127.pdf
[6] F. Alizadeh and R. Oprea, "Discovery and Mapping of the Dutch National Critical IP Infrastructure", M.Sc. thesis, University of Amsterdam, August 2013
[7] Hadoop Homepage, http://hadoop.apache.org/
[8] S. Madden, "From Databases to Big Data", IEEE Internet Computing, Volume 16, Issue 3, May-June 2012
[9] G.-J. Zwenne, "De verwaterde privacywet", lecture, Universiteit Leiden, April 2013, http://zwenneblog.weblog.leidenuniv.nl/files/2013/09/G-J.-Zwenne-De-verwaterde-privacywet-oratie-Leiden-12-apri-2013-NED.pdf
[10] Meldingsformulier verwerking persoonsgegevens, College Bescherming Persoonsgegevens, http://www.cbpweb.nl/Pages/ind_melden_formulier.aspx
[11] D.-E. Smørgrav, "Local caching resolver in FreeBSD 10", September 2013, http://blog.des.no/2013/09/local-caching-resolver-in-freebsd-10/
[12] DNSSEC Trigger Project, NLnet Labs, http://nlnetlabs.nl/projects/dnssec-trigger/
[13] S. Hollenbeck, "Extensible Provisioning Protocol (EPP)", RFC 5730, August 2009
[14] S. Krishnan and F. Monrose, "DNS prefetching and its privacy implications: when good things go bad," in USENIX Conference on Large-Scale Exploits and Emergent Threats (LEET), 2010, https://www.usenix.org/legacy/event/leet10/tech/full_papers/Krishnan.pdf
[15] Wet bescherming persoonsgegevens (WBP), http://www.cbpweb.nl/pages/ind_wetten_wbp.aspx
[16] Data Protection Directive, October 1995, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
[17] Data Protection Regulation, January 2012, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf
[18] SIDN Labs homepage, www.sidnlabs.nl
[19] AbuseHUB homepage, www.abuseinformationexchange.nl
[20] P. Mockapetris, "Domain Names – Concepts and Facilities", RFC 1034, November 1987
[21] P. Mockapetris, "Domain Names -- Implementation and Specification", RFC 1035, November 1987
[22] B. Moore, et al., "Policy Core Information Model—Version 1 Specification", IETF RFC3060, February 2001
[23] C. Hesselman, ".nl DNSSEC Deployment", DNSSEC Workshop at ICANN45, Toronto, Canada, October 2012, http://ccnso.icann.org/pt/node/34637
[24] Turris Project Homepage, https://www.turris.cz/en/
[25] Turris privacy policy, https://www.turris.cz/en/privacy
[26] M. Davids, "A Resolver Reputation System Based on Interpreting DNS Traffic Characteristics", 6th CENTR R&D Workshop, Paris, France, June 2014, https://centr.org/RD6-Davids-Resolver_Reputation_System
[27] P. Ohm, "Broken Promises of Privacy", http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006
[28] R. Leenes, "Acties tegen botnets door SURFnet en bij SURFnet aangesloten instellingen: privacy & data protectie aspecten", October 2013, http://www.surf.nl/binaries/content/assets/surf/nl/kennisbank/2013/expert_opinion_botnets_leenes_oktober_2013.pdf

APPENDIX A: POSSIBLE FILTERS

A filter is an operation performed on the personal elements of the DNS data (IP addresses or requested domain names), with a view to verifiably preventing disproportionate processing. One element of an ENTRADA privacy policy is a specification of the filters used for the ENTRADA application to which the policy applies (see also Section IV.B). In this appendix, we consider a number of filters that might possibly be applied.

A. Whitelisting

A whitelist may be compiled, specifying resolvers known not to belong to domestic users. Such resolvers would include those operated by, for example, ISPs, Google and OpenDNS. DNS queries that do not originate from a whitelisted resolver would then either be excluded from ENTRADA use, or used only after anonymisation. The approach could be employed either on the basis of static listing (i.e. using a list that changes only when a new resolver is detected) or on the basis of dynamic listing (i.e. using a list that may potentially include any resolver if it has sent a certain number of queries in a specified reference period).

Advantages of whitelisting: it provides assurance that recorded data cannot be used to compile a profile of a particular person. Disadvantages: (1) the detection of botnets with internal resolvers would be made more difficult or impossible, (2) the detection of botnets or other activities that make use of open resolvers would be made more difficult or impossible and (3) data could be overlooked if the list is not sufficiently up to date.

B. Blacklisting

A blacklist may be compiled, specifying the resolvers whose operators have indicated that data concerning queries from the resolvers should not be retained.

Advantages of blacklisting: users have the ability to control whether data regarding queries from their resolvers may be retained. Disadvantages: (1) the disadvantages of whitelisting apply equally to blacklisting, (2) the opt-out principle requires action by the user, which is inconvenient because the DNS is an infrastructure service that most end users are not familiar with, and (3) if ISPs choose to blacklist their resolvers, the ENTRADA platform will be denied a large amount of data, potentially negating its purpose (the enhancement of security and stability).

C. Non-anonymisation

Query data are not anonymised before being saved.

Advantages of non-anonymisation: no constraints are placed on the potential research uses of the data, thus maximising the scope for delivering secure and stable services. Disadvantages: (1) the data could potentially be used to compile a profile of a person who uses a private resolver, (2) the data must be secured to prevent access or processing by anyone except authorised internal personnel and (3) sharing the data with third parties would necessitate a separate anonymisation procedure or a confidentiality agreement.

D. Anonymisation of shared data

The filter anonymises data when they are to be shared with a third party.

Advantages of the anonymisation of shared data: the scope for research on the non-anonymised data is maximised and no constraints are placed on the potential internal uses of non-anonymised data. Disadvantages: (1) the effective anonymisation of data is no small matter and (2) it is feasible that the data lose their research value when anonymised.

E. Aggregation of addresses

It should be possible to add noise to the data, thus rendering it impossible to identify queries from particular IP addresses. Such a process would involve the aggregation of data to a higher level, so that individual queries are no longer traceable. The final x bits of a source IP address could be revised to zero. For example, zeroing the final eight bits of the IP address 192.0.2.189 would result in an address of 192.0.2.0.

Advantages of address aggregation: simple and quick to implement. Disadvantages: (1) follow-up of detected abuses would be impossible, because it would not be possible to ascertain the exact addresses of the offending machines and (2) research [27] has shown that, by utilising other attributes and external datasets, it is sometimes possible to trace data to a particular person even after aggregation.

F. Aggregation of addresses at autonomous network level

This form of filtering is similar to the aggregation of IP addresses, but, instead of elements of the addresses being changed to zeros, addresses are aggregated at the level of the network of origin. In many cases, the effect is the same if, for example, the final eight bits are deleted and a /24 network is involved. However, that is by no means always the case.

Advantages of this form of address aggregation: (1) it is a little more specific than general aggregation and (2) it is relatively straightforward to implement. Disadvantages: (1) it requires a network identification process and (2) loss of individual addresses can obstruct follow-up for some purposes.

G. General aggregation

General aggregation is similar to the aggregation of addresses (see Appendix A.E), but is a slightly more generalised process. Instead of aggregating the addresses, other data is aggregated. For example, the number of queries from each address can be counted and all other data aggregated.

The advantage of this form of aggregation is that it is easy to use. Disadvantages: (1) it is necessary to know exactly which data is required and (2) the way that this form of filtering would be used has not yet been precisely defined.

H. Distribution (distributed privacy preservation)

Distribution involves the partitioning of data across several entities (servers), so that no single entity has sufficient data to profile the behaviour of a particular address. With horizontal partitioning, queries remain intact but are distributed across several entities. With vertical partitioning, the various attributes of the queries are distributed across several entities.

The advantage of this type of filtering is that partitioning is relatively easy to realise through the use of multiple databases. Disadvantages: (1) if the separate databases become large enough, it can still be possible to compile profiles from the data stored on them; hence the number of queries stored in a single database must be capped, which can lead to a complex multiplicity of databases, and (2) partitioning will in practice be an obstacle to efficient analysis of the data, since the data must at some point be reunited; once the data are reunited, any privacy benefits secured by distribution may be lost.

I. Replacement

Attributes which could potentially be used to identify individuals can be replaced by dummy values. To enable analysis over a longer time period, the replacement must

always be one-to-one: an attribute value of x must always be replaced with the same value y.

Advantages of replacement: if the relationship between attribute x and replacement y is recorded, the original value of x can be subsequently ascertained if the need arises. Disadvantages: the value with which attribute x is replaced must be recorded, so that any subsequent occurrence of x can be replaced with the same value, thus complicating the parallel processing of new data.

J. Random modification

The modification of privacy-sensitive attributes by adding random data to them. IP addresses may be modified using a different method from that used for search terms. IP address 192.0.2.189 then becomes 203.0.113.044, for example.

Advantages of random modification: relatively straightforward to implement. Disadvantages: (1) it has not yet been established how difficult it is to de-randomise the data, (2) effective intervention following detection of abuse is not possible because it is not possible to ascertain the addresses associated with the abuse and (3) the scope for research is negatively influenced, because the identification of patterns on the basis of source and search term is almost impossible.

K. Omission of attributes

Omission of privacy-sensitive attributes such as source IP address and search term.

Advantages: relatively straightforward to implement. Disadvantages: (1) it has not yet been established how difficult it is to de-randomise the data, (2) effective intervention following detection of abuse is not possible because it is not possible to ascertain the addresses associated with the abuse and (3) the scope for research is negatively influenced, because the identification of patterns on the basis of source and search term is almost impossible.

Appendix B: Specimen policy

The Resolver Reputation R&D Policy will apply to our Resolver Reputation System [26], an ENTRADA application that assigns reputations to the resolvers that submit queries to the .nl name servers. So, for example, we may classify a resolver as 'suspect' if it appears to belong to a botnet. The aim is to forward the compiled information to the AbuseHUB, so that affiliated abuse desks can deal with infections. The Resolver Reputation System is currently under development at SIDN Labs and not in production use.

A. Identifier

Resolver Reputation R&D

B. Purpose

The purpose of the Resolver Reputation System is, as its name suggests, to assign reputation scores to the resolvers that look up .nl domain names. It is an experimental project, set up to establish whether we can use an automated system to detect the difference between a 'respectable' resolver and, for example, an infected machine that is trying to distribute spam.

C. Personal data

Because the system is concerned with specific machines, IP addresses are recorded. Otherwise, the main focus will be on information regarding the queries; the domain names specified in the queries will not be recorded, but the intention is to record whether the names contain more than two labels and what RR types the queries relate to. The number of queries will be recorded, as will the frequency with which particular header flags are set and the response codes.

The data for the last day will be retained. Once data become more than a day old, they will be aggregated to the last week, the last month, and the total. The times of the first and last queries are still specifically recorded, but otherwise it will not be possible to determine when a particular query was sent, unless the address has sent no more than two queries.

D. Filters

The filter that is used is 'General Aggregation' (see Appendix A.G). Individual queries from each IP address will not be saved, only total counts of certain properties of those queries. For example, 'from IP address 192.0.2.1, 50 queries containing the header flag TC were received'.

E. Retention

If a resolver has not been seen for 31 days, all data relating to that resolver is deleted from the system.

F. Access

For the time being, access will be limited to internal staff within the discrete lab environment at SIDN Labs. Access will be controlled by the use of passwords and client certificates. If the system is upgraded to a production service, this policy must be updated and re-evaluated.

If we share data with any third parties, either the third parties in question will be restricted to the owners of the networks to which the investigated addresses belong, or the data that are shared will be restricted to totals, devoid of individual IP addresses.

G. Type

This policy applies to the R&D phase of the Resolver Reputation Project.

H. Other security measures

No other security measures are applicable.

DOCUMENT HISTORY

Version	Date	Major changes
1.3	30-Sep-2014	First public version
1.4	4-Nov-2014	Added application silos, clarified distinction between R&D and production