Internet Crime Forum

Principal Current Data Types

Howard Lamb
Chair
ICF Data Retention Project Group

March 2003

NOTICE OF COPYRIGHT AND LIABILITY

Liability

Whilst every care has been taken in the preparation and publication of this document, ICF, nor any committee acting on behalf of ICF, nor any member of any of those committees, nor the companies they represent, nor any person contributing to the contents of this document (together the “Generators”) accepts liability for any loss, which may arise from reliance on the information contained in this document or any errors or omissions, typographical or otherwise in the contents.

Nothing in this document constitutes advice. Nor does the transmission, downloading or sending of this document create any contractual relationship. In particular no licence is granted under any intellectual property right (including trade and service mark rights) save for the above licence to copy, store and distribute this document and to produce derived copyright works.

The liability and responsibility for implementations based on this document rests with the implementer, and not with any of the Generators. If you implement any of the contents of this document, you agree to indemnify and hold harmless the Generators in any jurisdiction against any claims and legal proceedings alleging that the use of the contents by you or on your behalf infringes any legal right of any of the Generators or any third party.

None of the Generators accepts any liability whatsoever for any direct, indirect or consequential loss or damage arising in any way from any use of or reliance on the contents of this document for any purpose.

If you have any comments concerning the accuracy of the contents of this document, please write to: secretariat@internetcrimeforum.org.uk

Contents

1. Introduction

1.1 In December 2001, the Internet Crime Forum (ICF) established a project group, the primary aim of which was to identify which data types are currently associated with subscribers who have access to the Internet.

1.2 The group was not tasked with debating the legal issues in relation to the data types identified. There are many legal issues relating to data retention and these will undoubtedly be discussed in other documents

1.3 The group was established with a view to producing a document that would provide a better understanding of the technology used and the information that law enforcement is seeking to assist its investigations.

1.4 It is not intended to be a standard or a best practice document. The document is intended to be a guide to what data types may be available to law enforcement when conducting an investigation. It does not recommend or guarantee what data types may be available, or for how long each data type might be retained (if it is logged at all).

2. Group Members

2.1 The group is restricted to technical and investigation experts, as explained in 1.2 this group does not hold a view on the value or legality of access to this data.

2.2 The ICF Data Retention Project Group called upon experts from the Internet industry who gave advice on the numerous data types that are created when a subscriber connects to and communicates via the Internet. This connection could be through an Internet Service Provider (ISP), a Virtual Internet Service Provider (VISP) or by other connection to the Internet

2.3 The group also engaged the services of Computer Forensic experts whose work regularly involves liaising with various Law Enforcement Agencies and assisting with their investigations involving the Internet. Representatives of various Trade Associations were involved in the process, together with several members of various Law Enforcement Agencies.

3. Acknowledgements

We would like to acknowledge the support given to this project by Chief Superintendent Len Hynds of the National High Tech Crime Unit, members of the Internet Crime Forum, and to those experts from the Internet and Forensics Industry who have assisted in the process. All participants gave freely of their time as they agreed it was vital that this type of work be carried out.

4. Current Data Types

4.1 This document seeks to identify the principal known Data Types that a subscriber to an Internet Service might create whilst they are actively subscribing to and utilising their Internet account.

4.2 It is accepted that this document could not be a definitive document of all data types due to the rapid development of technology.

5. Service Providers

5.1 **It must be appreciated by the reader of this document that not all Internet Service Providers retain the data types that are mentioned within this document.**

5.2 Each service provider is aware of their current data retention practices and may be able to advise on the detail. Communication should in the first instance be routed through a Law Enforcement Single Point of Contact for Law Enforcement personnel. Requests for data retention policies made from outside the SPOC regime may be liable to conditions determined by individual ISPs.

5.3 There are service providers, known as Virtual Internet Service Providers (VISPs), who utilise most, if not all, the infrastructure of a large service provider. They may combine various elements of a service, such as e-mail, sign up servers, radius servers, web cache and usenet news and badge them as their own. In these cases the data that a subscriber generates may be spread across several companies

5.4 Even amongst traditional ISPs some parts of their service may be provided by third parties. In this case as well, information may be held by many different companies and may or may not be accessible to the primary ISP.

5.5 Furthermore, some data types for example, web server log information, may be owned by and under the control of the customer rather than the ISP.

6. Glossary

There is a glossary attached to this document that informs the reader of what the various data types are and it is advisable that this is read in conjunction with the rest of the document.

7. Subscriber Data Types

7.1 The attached diagram identifies the principal data types that may be created when a subscriber accesses the Internet.

7.2 The data types have been broken down into two main areas. The first being the activity of the subscriber and the second the resources that a subscriber could utilise.

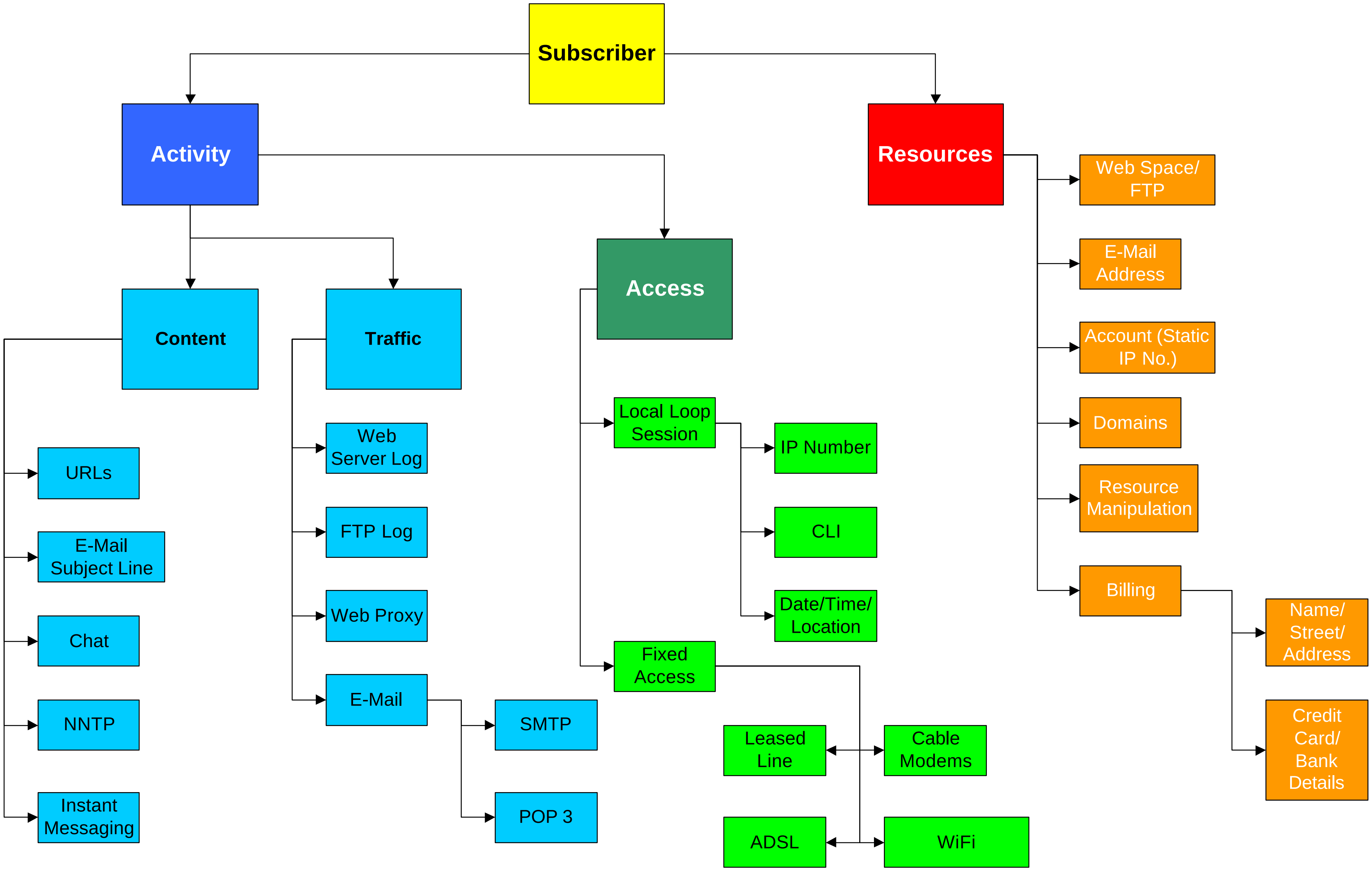
7.3 When matching events on the Internet with details recorded in ISP logs it is absolutely essential to ensure that time and date information is correctly recorded. It is Best Practice for ISPs to synchronize their systems with global time standards using protocols such as NTP, however consideration should always be given to this not being the case for particular logs. Equally it is essential that enquiries about logging information provide accurate timing information. A frequently encountered pitfall is incorrect handling of timezone offset information and careful attention should be paid to this.

7.4 This is not a definitive list of data types and it must be appreciated that advances in technology may well mean that some of the data types that are currently of little value may at some stage in the future generate logs that could be useful for the purposes of investigations.

7.5 The table below identifies each of the data types and the data that could be generated by the subscriber.

7.6 Data can only be obtained in accordance with UK Law and international treaties. This document does not address this issue any further.

7.7 Internet Service Providers retain data for business purposes. The procedures surrounding this data retention may affect the way in which data could be used for evidential purposes.


Subscriber
Activity
Resources
Access
Content
Traffic
URLs
E-Mail Subject Line
Chat
NNTP
Instant Messaging
Web Server Log
FTP Log
Web Proxy
E-Mail
SMTP
POP 3
Local Loop Session
IP Number
CLI
Date/Time/ Location
Fixed Access
Leased Line
Cable Modems
ADSL
WiFi
Web Space/ FTP
E-Mail Address
Account (Static IP No.)
Domains
Resource Manipulation
Billing
Name/ Street/ Address
Credit Card/ Bank Details

Activity	Data Type	
Content		
URLs		A URL (Uniform/Unique Resource Locator) is the address of a file (resource) accessible on the Internet. The type of resource depends on the Internet application protocol. Using the World Wide Web's protocol, known as HyperText Transfer Protocol (HTTP), the resource can be an HTML page, an image file, a program such as a common gateway interface (CGI) application or Java applet, or any other file supported by HTTP. The URL contains the name of the protocol required to access the resource, a host name that identifies a specific computer on the Internet, and a hierarchical description of a file location on the computer. The host name can be used to determine the physical location of the computer and its logical ownership.
E-mail		ISPs may hold e-mail on behalf of subscribers. Much of the e-mail is content and a number of different legal regimes apply to the divulgence of this. Some of the header is communications data. In addition, details of what e-mail has been sent and received may be recorded in logs. Some of the information in these logs may be content.
Chat		Depending upon the technology, the service provider will not normally retain the content of an individual Chat Room session but individual participants will be able to make their own record. It may be possible to trace and identify participants in a chat session provided the IP address, or for some ISPs the screen name, is obtained together with an accurate time stamp.
NNTP/Usenet		In order to trace the author of a Usenet article, the article headers will need to be inspected. These will usually contain the posting IP address and time stamp. The system through which it was originally posted should then be able to identify the account responsible for creating the posting. Provision of Usenet services is increasingly performed by third parties so it may be necessary to make further enquiries with a connectivity ISP to determine where the account was used from. The content of an NNTP (Usenet) session will not be retained by a service provider. Therefore the readership of an article is unlikely to be available. Usenet postings are intended to be exchanged between ISPs. This means that an article will often have have been posted on a different service provider from the one on which it is read.
Instant Messaging		Instant messaging (sometimes called IM or IMing) is the ability to easily see whether a chosen friend or co-worker is connected to the Internet and, if they are, to exchange messages with them. Instant messaging differs from ordinary e-mail in the immediacy of the message exchange and also makes a continued exchange simpler than sending e-mail back and forth. Most exchanges are text-only. However, some services allow attachments. In order for IMing to work, both users (who must subscribe to the service) must be online at the same time, and the intended recipient must be willing to accept instant messages. (It is possible to set your software to reject messages.) An attempt to send an IM to someone who is not online, or who is not willing to accept IMs, will result in notification that the transmission cannot be completed. The ISPs will not in general have any records of the messages which have been exchanged because they flow directly between the participants (Peer to Peer). If a rendez-vous server is involved in the initial connection between the participants then some logging information about their identities may be retained. The rendez- vous server may be totally independent of any connectivity ISP.

Activity	**Data Type**	
Traffic		
Web Server Logs		These typically contain the source IP address, requested content, submitted data e.g. username, password and previous site visited. Some of the data may be content rather than traffic data. Some of the data may be anonymised in near real time. Some of the IP addresses may be proxy caches rather than the actual requestor.
FTP Logs		These contain source IP address, account details and details of the file names uploaded into or downloaded from. Although most sites appear to have a username/password login, anonymous guest accounts are also common and although an e-mail address is traditionally provided as identification there is seldom any validation of this whatsoever. Some of the data may be content rather than traffic data. It is quite common for customers to upload the content of their web pages using FTP.
Web Proxy		A proxy server is a server that acts as an intermediary between a user and the Internet. A proxy server receives a request for an Internet service (such as a Web page request) from a user. If it passes filtering requirements, the proxy server will access the remote site and pass the information to the user. A web cache maintains a store of previously downloaded items from the Web such as an HTML page. If it is asked for a page that is already in its store, then it returns it to the user without needing to forward the request to the Internet, though it may need to check if its cached copy remains up-to-date. If the page is not in the cache then the cache server acting as a client, on behalf of the user, uses one of its own IP addresses to request the page from the server out on the Internet. The user's general impression of using both proxy servers and caches will be of a direct connection to the remote site. In the ISP context it is usual to combine these two functions, and the result may be, rather confusingly, called a web cache, a web proxy or indeed a proxy-cache. Some ISPs use a "transparent" scheme that intercepts, for example, all HTTP (port 80) traffic and sends it via a proxy-cache. In other cases the use of a proxy-cache is entirely under the users' control, though the ISP may encourage usage by means of the default configurations shipped to its customers. These servers can produce logs of the data handled, giving the local customer IP address, details of the requested content and details of any connections made to remote sites. Complete logging may only be enabled for troubleshooting, but even incomplete logging can create very substantial volumes of data and these logs are not kept for long periods of time. The presence of a proxy may mean that the user never accesses the target web site. The access will show the IP address of the proxy server in the Web log. The proxy server may be configured to pass information to the target web site giving some details of the user, but some servers are configured specifically to obscure the true identity of the user. Some web pages are designed so that they cannot be cached and so this traffic will flow directly and hence proxy-cache logs will be incomplete. Similar effects will be caused by the use of protocols such as HTTPS which often avoid the use of a proxy-caches altogether.

Activity	Data Type	
Email		
		SMTP (Simple Mail Transfer Protocol) is used for sending and receiving e-mail between permanently connected machines. However, because mail is "pushed" rather than "pulled" it works very poorly with intermittently connected machines. Therefore it is usual to provide mail delivery to dialup customers via POP3 or IMAP. These provide a store and forward system, so that users can periodically "pull" any new e-mail. SMTP is the standard method for ISP customers to send their e-mail. Although a few user systems can be made to deliver direct to remote systems, it is more common to relay e-mail traffic via an SMTP server at the ISP called a "smart host". Some ISPs will intercept all outgoing SMTP (port 25) traffic and force it to use the smart host. POP3 is a relatively simple protocol for e-mail reception. It is usual to configure clients to delete the e-mail once it has been fetched. If it is not deleted then the ISP will delete it after a preset period. Long term storage is done on the client machine. IMAP is somewhat more complex and provides a client/server implementation of a fully featured e-mail interface with all the e-mail held on the server machine, possibly for very long periods. Many IMAP clients can also be configured to hold a copy of the email content locally.
SMTP		Mail will be held on an SMTP server until it can be passed to a destination, but most service providers do not routinely keep content thereafter. A service provider may retain summary logging details of e-mail that has been received from or sent to their customers. This would include a unique message identifier, who the mail was alleged to be from, who the mail was addressed to, the IP address of the immediately previous hop and the time and date the mail was sent. Further information such as size and content such as the subject line may sometimes be recorded. When the intended destination of e-mail is unavailable it may be routed via intermediate machines (using lower priority MX records). This will reduce the usefulness of IP address logging details. It is also essential to view the "from" details with caution since they are trivial to forge. Finally, many ISPs have outsourced virus scanning and "spam" deletion services. Initial delivery is made to a third party who will only forward genuine e-mail to the ISP's systems. In all cases, the e-mail itself should contain full details of all the machines it has passed through, but each machine will almost invariably only record one part of its journey.
IMAP & POP3		IMAP & POP3 logs typically contain just brief summary details of connections. These may extend as far as recording the connecting IP address and how many e-mails were read or deleted. It would be very unusual indeed to record anything which references the content, sender or path associated with transmission of the e-mail. IMAP & POP3 servers can usually be accessed from anywhere on the Internet so any IP address recorded may well require further tracing to be useful.
Webmail		Access to e-mail via a web interface may be provided as a front end to POP3 services or as a service in its own right. Logging will typically record the IP address that accesses the mail box and may record which items of mail were looked at. Webmail services are almost invariably designed to be used from any Internet address.

Activity	Data Type	
Access		
IP Address		For both circuit switched and fixed services, an IP address can either be static (allocated on permanent basis) or dynamic (a different IP address allocated each time authentication is made, or reviewed after a fixed amount of time). When mapping a dynamic address to an account it is therefore essential to provide accurate timing information (date, time and timezone).
Account usage		Logging of account usage will record the date and time that the connection was established, and the date and time that it ceased. Further details such as the number of packets transferred may also be available from some ISPs. Dial-up account authorisation if often done using a system called RADIUS so these logs are often referred to as RADIUS logs. A number of different versions of RADIUS are in use, so that the actual format of the logs (and indeed the format of the time and date information) may vary from ISP to ISP. Further logs, holding much the same information, from the Network Access Servers (NASs), may also be available at some ISPs. Many fixed services are shared between a number of local users and sometimes these users will have their own fixed IP addresses. Sometimes they may have (even if the fixed service has a static IP address) a dynamic IP address allocated by their local system administrator, who may have DHCP logs available.
Circuit Switched		
		Dial-up services such as POTS, ISDN, GSM, GPRS where each Internet connectivity session is established on demand. These are sometimes provided "free" with the ISP gaining revenue from the interconnection payments within the wholesale telephone network. In such instances there may not be verified billing/subscriber details. The subscriber may use the service from multiple locations, or move to another permanent address without the ISP becoming aware of it.
CLI		If CLI is captured by the service provider it is most likely to be recorded within the RADIUS logs. Some ISPs will require non-withheld CLI to be presented to the ISP, but others don't. Some types of account will require access only from an authorised CLI. At present, ISP equipment will not usually record the CLI if it is marked by the caller to be withheld. If the subscriber is calling from within an organisation, then the CLI will sometimes only identify the organisation, not the subscriber (and his extension number).
Fixed Access		
		Unlike circuit switched services which may be provided "free", almost all fixed access systems are charged for. This means that a valid billing address will be present in the ISP's accounting systems. In addition an installation address will be recorded, though in some cases the service may be moved to another address without the ISP becoming aware of it or bothering to record the change.
Leased Line		A leased line is a dedicated link via the local telephone exchange that has been provided for private dedicated use. A leased line is usually contrasted with a circuit switched or dial-up connection.
Cable modem		A broadband connection delivered using technology associated with cable-tv. There is usually a dedicated cable modem linked from the television cable.
ADSL		A method of providing broadband access over a standard telephone connection. Within the UK this is mainly provided by British Telecom who route the traffic over a high speed ATM backbone and provide an IP data connection via various branded ISPs who add their own email and web services.
Satellite		A system designed primarily for rural areas to provide often slow speed broadband Internet access. The return path from the user to the ISP may be via dial-up or via the satellite.

Activity	Data Type	
Access		
	Wifi	A fixed base station is connected to the Internet by either dial-up or fixed connection. Clients may access by wifi standard within a limited distance. Currently these systems are insecure even where encryption has been used to protect the connection. Logs may only show that someone (necessarily physically close to the base station) has been using the system but local logging may provide further traceability. Some WiFi installations are deliberately made open for public access and some commercial operations provide access for payment, which may be made in cash.
Resources		
	Shell Sessions	Details pertaining to telnet and other 'shell' login sessions may be held in several files (telnet connections are typically logged in 'last' and 'messages' files on UNIX based systems). Shell sessions may log a variety of data including start/stop and source IP address.
	Web/FTP Space	Web and FTP Space may be provided separately or as part of a service package. All of the remarks relating to billing (to identify the owner) to server logs (to identify readers) and to FTP Logs (to identify up loaders) apply to this section.
	E-Mail Addresses	There is a mapping between the e-mail address and the account. This may vary between ISPs. An account may have one or more e-mail addresses associated with it. Users may have the ability to change e-mail addresses at will. Some ISPs may not hold data on previous e-mail addresses.
	Domains	ISPs provide Domain Name Service (DNS) to allow mapping between domain names and IP addresses, as well as information such as where to deliver e-mail. Details of who actually owns the domain name will be held by the appropriate registrar. The ISP will have some records as to which of their customers is controlling it. There may be limited records of historical settings.
	Resource Manipulation	Many services provided by ISPs, and particularly those provided by third parties, can be configured by the user. For example, e-mail may be re-directed to another account, web space requests may be directed to another server, or DNS settings may be rearranged. The system that is used for this configuration may keep logs that allow historic configurations to be reconstructed.
Billing		
		Many forms of access are paid for. Billing data may relate to an individual or could also be that of an organization. Some systems may be sub-let and billing records will relate to the 'letting company'. Many services are re-sold. **Some systems may be insecure and used without permission.**
	Name, Street, Address	A service provider does not necessarily verify a subscriber's name and address details. This is dependant upon the service a subscriber utilises whilst on the Internet. In many instances the subscriber will provide CLI (Caller Line Identifier) as a part of the authentication process prior to their use of that service. This CLI can often be mapped to a geographical location by the appropriate telco.
	Credit Card/ Bank Details	Accounts where payments are made, credit card, debit card, direct debit, cheques or standing order will provide traceability through the banking system. Where postal orders or cash payments are made or accepted, these will not always be verified. It should be noted that billing information may not be retained by the backbone ISP but by the Virtual ISP who has ownership of the customer.

Glossary of Terms used Within this Document

Access	Data access is being able to get to (usually having permission to use) particular data on a computer.
ADSL	Asymmetric Digital Subscriber Line is a technology for transmitting digital information at a high bandwidth on existing phone lines.
ATM	Asynchronous Transfer Mode. A switching technology for transferring packets of data. ATM was originally developed for voice application, but is now used for Internet transports and underpins current broadband technologies.
Broadband	A High Speed always-on Internet connection. Normally at a speed of between 128Kbps and 2Mbps
Cable Modem	A cable modem is a device that enables you to hook up your PC to a local cable TV connection in order to send and receive data packets.
CGI	Common Gateway Interface. A method of providing dynamic content within "web pages".
Chat	Facility to talk with others whilst on line.
CLI	Calling Line Identifier. The telephone number of the local loop that a person has used to access a dial-up service.
DHCP	Dynamic Host Configuration Protocol. A protocol that lets network administrators automate and manage centrally and the assignment of IP addresses in an organization's network.
DNS	Domain Name System. A protocol for providing mappings from domain names to resource identifiers such as IP addresses.
Domain name	A domain name is a user-friendly method of identifying the location of resources on the Internet.
E-mail	E-mail is the exchange of electronic messages using telecommunications systems.
E-mail Subject Line	A conventional e-mail header that is intended to provide a brief description of the contents of an e-mail.
FTP	File Transfer Protocol. A protocol for transferring files between machines. FTP is often used to upload the content of web sites onto servers.
GPRS	General Packet Radio System. An always-on data service built on GSM. Sometimes known as 2.5 generation, to distinguish it from 3rd Generation digital phones.
GSM	Global System for Mobile Communication. Second generation circuit-switched digital mobile telephony and data system.
HTML	Hypertext Markup Language. This is the language which is used for "web pages" to indicate the structure of documents so that browsers can display them in a standardised manner.
HTTP	Hypertext Transport Protocol. A protocol for transferring "web pages" from one machine to another.
HTTPS	Secure HTTP. Transfer of "web pages" over an encrypted transport protocol.
ICF	Internet Crime Forum. A body formed "To promote, maintain and enhance an effective working relationship between industry and law enforcement to tackle crime and foster business and public confidence in the use of the Internet in ways that respect human rights and are sympathetic to the needs of industry."
Instant Messaging (IM)	A quick and easy way of exchanging messages with others who are also online.
IMAP	Internet Message Access Protocol. A protocol for accessing e-mail that is received, organised and stored on a remote server.
MX record	DNS record entry that indicates where e-mail for a domain is to be delivered.

IP	Internet Protocol. The basic protocol used for communication between computers on the Internet.
IP address	Internet Protocol Address. A numeric value that serves to uniquely identify an interface that is connected to the Internet. Most computers connected to the Internet have just one relevant interface, and "IP address" is therefore often used as shorthand for the address of a machine connected to the Internet.
ISDN	Integrated Services Digital Network. A digital circuit-switched telephony and data service.
ISP	Internet Service Provider. An organisation that provides Internet services to its customers. ISPs provide connectivity to the Internet, along with many other services such as e-mail and web space. However, some customers may use a number of different ISPs to cover different aspects of their requirement.
Java applet	A way of providing "mobile code" on web pages so as to enable extra functionality on web pages.
Leased Line	A method of providing a fixed connection to the Internet.
Local Loop	In telephony, a local loop is the wired connection from a telephone company's telephone exchange its customers' telephones at homes and businesses
NAS	Network Access Server. The "modems banks" used by an ISP to receive POTS and ISDN calls from their subscribers and transfer data to the Internet.
NNTP	Network News Transfer Protocol. A protocol for transferring Usenet articles over the Internet.
NTP	Network Time Protocol. A protocol that is used to synchronize computer clock times in a network of computers, such as the Internet.
POP3	Post Office Protocol 3. A protocol for collecting e-mail from a server.
POTS	Plain Old Telephone System. The traditional analogue circuit switched telephony service.
PSTN	Public Switched Telephone Network. Often used as a synonym for POTS, but also includes ISDN.
RADIUS	Remote Authentication Dial-in User Service. A protocol for communicating authentication information and establishing parameters for dial-up connections to the Internet.
SMTP	Simple Mail Transfer Protocol. A protocol used for the transport of e-mail over the Internet.
SPOC	Single Point of Contact. A scheme whereby requests from law enforcement organisations are funnelled through a single part of that organisation and passed to a single contact point within ISPs.
TCP	Transport Control Protocol. A protocol layered over IP that provides a reliable delivery service for data.
UNIX	A computer operating system widely used by ISPs.
URL	Unique/Uniform Resource Locator. A stylised naming system for web resources.
VISP	Virtual ISP. An ISP whose infrastructure is completely provided by third parties.
Web Cache	A cache is a server that retains copies of web content so as to provide timely local delivery for repeat requests.
Web Proxy	A proxy is a server that acts as an intermediary between a workstation user and the Internet
WiFi	Wireless systems (such as 802.11) that provide Internet connectivity.