Unterschrift Doz. Dr. Wilfried Gansterer

TECHNISCHE UNIVERSITÄT WIEN

VIENNA UNIVERSITY OF TECHNOLOGY

MASTERARBEIT

Implementation & Consequences of the EU Directive 2006/24/EC in the Context of Internet Access and E-Mail

Ausgeführt am Institut

Research Lab Computational Technologies and Applications

der Universität Wien

unter der Anleitung von

Doz. Dr. Wilfried Gansterer

durch

Gerald Stampfel

Dampfschiffstraße 12/1/2
A-1030 Wien

Wien, am 6. Februar 2008

Eidesstaatliche Erklärung

Ich erkläre an Eides statt, daß ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 6. Februar 2008 Gerald STAMPFEL

Abstract

The Data Retention directive 2006/24/EC of the European Parliament, released on 15.03.2006, requires the operators of publicly accessible electronic communication networks to store and provide traffic and location data generated or processed in their networks to serve the investigation, detection, and prosecution of serious crime. This thesis focuses on the technical and financial issues related to the implementation of the directive with respect to the guidelines for Internet access and Internet e-mail. Instead of each Internet Service Provider (ISP) having to implement the guidelines for both areas, the assumption made here is that service providers offering exclusively Internet access or Internet e-mail services may also be obliged to implement only one of the two areas. In order to illustrate the conclusions drawn in this thesis, cost estimations are made for a fictitious medium-sized Austrian service provider.

Kurzfassung

Die EU Richtlinie 2006/24/EC vom 15.03.2006 schreibt den Betreibern öffentlicher Kommunikationsnetze vor, diverse Verkehrs- und Standort-Daten, die von ihnen erzeugt oder verarbeitet werden, auf Vorrat zu speichern, um sie auf Anfrage zur Ermittlung, Feststellung und Verfolgung von schweren Straftaten den Behörden zur Verfügung zu stellen. Diese Arbeit setzt sich mit den technischen und finanziellen Konsequenzen einer Implementierung der EU-Richtlinie für die Anbieter von Internetzugangs- und Internet-E-Mail Diensten auseinander. Der Ansatz, der hier gewählt wurde, ist der, dass ein Dienste-Anbieter nur die Teile der Richtlinie erfüllen muss, die für ihn zutreffen, er also auch entweder nur Internetzugangs-Anbieter oder nur E-Mail Dienstbetreiber sein kann. Um die Ergebnisse dieser Arbeit zu veranschaulichen, wurden die finanziellen Kosten und der benötigte Speicherplatz für einen fiktiven mittleren österreichischen Dienste-Anbieter geschätzt.

Dank ...

... *meiner Familie*, speziell meinen *Eltern*, für Ihre unentbehrliche Unterstützung seit dem Tag 0

... *Marie* für die Liebe

... *Benna*, *Chriz*, *Evil Hias*, *Korl*, *Norb*, *Surfinstructor* und *Thomaso Buscetti* für die Freundschaft

... *3 Feet Smaller, Arctic Monkeys, Art Brut, Audioslave, Babyshambles, Beatsteaks, Billy Talent, Blink 182, Bloc Party, Blur, Dandy Warhols, Die Ärzte, Eagles of Death Metal, Foo Fighters, Franz Ferdinand, Goldfinger, Greenday, Heinz, Hot Hot Heat, Iggy Pop, Incubus, Jet, Jimmy Eat World, Julia, Juliette & the Licks, Kaiser Chiefs, Kings of Leon, Lenny Kravitz, Lost Prophets, Mando Diao, Maximo Park, Moldy Peaches, Muse, Nada Surf, Offspring, Ok go, Pearl Jam, Phantom Planet, Pink, Powderfinger, Puddle of Mud, Queens of the Stone Age, Red Hot Chili Peppers, Sebadoh, Slut, Sophia, Sportfreunde Stiller, Sum41, The Ataris, The Caesars, The Darkness, The Distillers, The Donnas, The Fratellis, The Hives, The Killers, The Kills, The Kooks, The Libertines, The Rakes, The Sounds, The Soundtrack of our Lives, The Staggers, The Strokes, The Subways, The Vines, The White Stripes, The Yeah Yeah Yeahs, Tomte, TV On the Radio, We are Scientists, Weezer, Wir sind Helden* und *Wolfmother* für die Musik.

Außerdem bedanke ich mich sehr bei meinem Betreuer *Dr. Wilfried Gansterer* für seine stets simplen und genialen Lösungsvorschläge zu den Problemen die sich im Rahmen dieser Arbeit ergaben.

Contents

1 Introduction

In this document, the implications of the EU Data Retention directive 2006/24/EC and its implementation for the areas Internet access and Internet e-mail are analyzed. The basis and reference point for all investigations summarized in this thesis is the original formulation of the Data Retention directive [11]. No national legal regulations derived from this directive were taken into account.

The structure of this thesis is as follows: **Chapter 1** summarizes the motivation for this work and a study carried out in the Netherlands, **Chapter 2** reviews some technical background and foundations, and **Chapters 3 and 4** are the two central chapters discussing the data retention of Internet access and e-mail. **Chapter 5** summarizes this thesis.

1.1 EU Directive 2006/24/EC

The Data Retention Directive 2006/24/EC of the European Parliament, released on 15.03.2006, orders the operators of publicly accessible electronic communication networks to store certain data that is generated or processed in their networks to serve the investigation, detection, and prosecution of serious crime. The national service providers are ordered to implement and maintain the technical means required to store and provide this data to government authorities.

For each of the three categories of public network operators, namely Internet access, Internet e-mail, and Internet telephony operators, is defined which data has to be retained. Affected are traffic and location data, for a time period of between six months and two years. Although the directive also orders mobile and fixed telephony network operators to store traffic and location data, the *focus of this thesis is on Internet access and Internet e-mail only*. The thesis at hand tries to analyze what can be done, with the current technical means available, to fulfill the requirements of the EU directive and how it can be done.

1.2 Related Investigation in the Netherlands

The goal of the Dutch report [58] was to identify the organizational and technical changes and costs for the providers and requesters caused by an implementation of the EU directive 2006/24/EC. The report was requested by the Dutch government and carried out by the consulting com-

pany Verdonck, Klooster & Associates (VKA)[1] in cooperation with Lucent Technologies[2], an international vendor of telecommunications equipment. They worked in close association with the national providers of telecommunication services and delivered the report in September 2006.

There is already an ISP-to-government interface implemented in the Dutch telecommunication environment, the Centraal Informatiepunt Opsporing Telecommunicatie (CIOT) which is a part of the Dutch Ministry of Justice. Providers of mobile and land line communications and of Internet access in the Netherlands have to provide location data including name and address, IP and e-mail addresses, and login names to the CIOT on a daily basis. This information can be accessed by authorized criminal investigators from their offices via a secured line. According to a Dutch Newspaper, 1.2 million of such requests have been made in 2004 to the CIOT [36].

The investigations of VKA started with working out a set of distinct implementation options. These have been identified by varying three distinguishing aspects:

Centralized storage vs. decentralized storage: This refers to the national context of the information storage. Either every provider holds his own data or the data is transferred to a nationwide store like the CIOT which is further referred to as an intermediary third party in the context of centralized storage.

Response by the provider vs. direct access by the requester: The access to the retained data by government authorities may be either direct or human-mediated: With standardized requests and responses and clearly defined interfaces, the process of accessing a provider's data store can be designed to not depend on human action on the provider side. The alternative is a manual request response option which is carried out by an employee of the provider.

Correlated storage vs. uncorrelated storage: The correlation here describes if the traffic and location data at the providers' side are stored as logically connected records. Location data refers to personal details about the individual of interest such as name, address, etc. Traffic data refers to information concerning actual communication like e-mail messages or Internet telephony conversations. Correlated storage would enable the requester to ask for the location data and the traffic data in one request whereas uncorrelated storage would need him to ask twice in order to gather all the information.

Grouping of these variables leads to six different implementation options. Two additional models have been added to the investigation: "Hybrid

[1] http://www.vka.nl
[2] http://www.alcatel-lucent.com

storage, correlated data" and "hybrid storage, uncorrelated data". With this option, the traffic data is stored at the providers, but the location data is transferred to a nationwide third party. The idea behind this is that the requester may ask the intermediary third party in a first step which provider is serving the customer he is searching for (location data). In a second step, he asks the returned provider for the demanded telecommunication activities of this specific customer (traffic data). Table 1.1 summarizes the implementation options investigated.

	Correlated data storage	**Uncorrelated data storage**
Data storage location and information access	Decentralized storage, response by provider	Decentralized storage, response by provider
	Decentralized storage, direct access	Decentralized storage, direct access
	Centralized storage, direct access	Centralized storage, direct access
	Hybrid storage, direct access	Hybrid storage, direct access

Table 1.1: Qualitative assessment: eight implementation options

The authors of [58] further proceeded by gathering opinions on these eight implementation options working in close association with the Dutch ISPs. An assessment model was developed in order to compare them systematically. The assessment can be roughly divided into two aspects: A quantitative assessment providing a cost estimation on the national level and a qualitative assessment weighting the strengths and weaknesses of each model. The assessments were based on analyzing the results of written questionnaires handed to the providers.

1.2.1 Quantitative Assessment Model

The quantitative assessment model investigates the costs associated with the implementation of each option. VKA identified the core processes *acquisition*, *storage*, and *retrieval* and tried to quantify the expenses of each for the particular actors. Four of the implementation options involve only two actors, the provider and the requester, whereas the two hybrid and the two centralized options add a third actor: The intermediary third party.

In order to estimate the costs, a few basic assumptions about a typical Dutch provider had to be made. A fictitious provider was introduced with 125 000 clients and one fixed telephone account, one mobile telephone account, one Internet access account, and one Internet e-mail account per client. Such a company represents a medium-sized telecommunication services provider in the Dutch situation. By assuming average behav-

ior patterns obtained from Statistics Netherlands[3] for the customers and combining them with the costs of the various technical and organizational aspects, the costs of this medium-sized provider were estimated. A nationwide cost estimation was done on the basis of an extrapolation to five million customers for big providers and to 1 000 customers for small providers. Scale advantages of 80% of the extrapolated costs were assumed for big providers and scale disadvantages of 200% were assumed for small providers.

The number of requests were estimated to increase by 20% each year and the data retention period was assumed to be one year.

1.2.2 Qualitative Assessment Model

Together with an advisory committee composed of representatives of government and telecommunication providers, 100 points were distributed among five categories: Organization and processes (30 points), technology (30 points), business case (10 points), information security (20 points), and implementation term[4] (10 points). The eight implementation models were evaluated with respect to these categories. The particular number of points provided above is the maximum score that each implementation option may reach in this assessment. This implies that any implementation option could score at most 100 points in this qualitative assessment. Using the providers' answers to the questionnaires, a score per category and a total score for each option were calculated.

1.2.3 Results

Each of the implementation options was examined with respect to the features provided and the associated implementation and operation expenses. This section summarizes and discusses the results found in [58].

As Table 1.2 shows, in the qualitative assessment the option centralized storage, direct access scored highest. The option decentralized storage, response by provider got the most points for the categories business case and implementation term because new services can be introduced very fast and there is no coordination required with other parties. This option also saves information technology resources and time needed for implementation because there is no need to create interfaces for requests, standardize responses, and implement direct access to the retained data. The same option scores lowest regarding the information security because the requested data is processed by humans. The highest score in terms of information security is reached by the options with the highest degree of automation: Centralized storage, direct access and hybrid storage, direct access. Centralized storage, direct access scores highest for organization and processes

[3] http://www.cbs.nl

[4] "Implementation term" refers to implementation feasibility

too, resulting from the efficiency of the query process and the possibility to automatically prepare reports to the European Commission.

	Decentralized storage, response by provider	**Decentralized storage, direct access**	**Centralized storage, direct access**	**Hybrid storage, direct access**
Organization and processes	12	9	18	18
Technology	18	10.2	20.1	20.1
Business case	7.5	0	2.5	2.5
Information security	4	11	20	15
Implementation term	10	0	0	0
Total	51.5	30.2	60.6	55.6

Table 1.2: Results qualitative assessment

The results of the quantitative assessment are shown in Table 1.3. These numbers are extrapolated from the model provider with 125 000 clients to the entire market in the Netherlands and therefore to nationwide expenses. The costs for the model provider employing centralized storage with direct access are estimated to be about €206 500 in the first year for a data retention period of one year. This is equivalent to costs of roughly €0.14 per subscriber per month.

When comparing the correlated version of a single implementation option with its uncorrelated version in Table 1.3, it is interesting to observe that the latter is always a bit more expensive than the former for most of the cases. The higher costs for uncorrelated storage are caused by higher operational costs which result from more requests compared to correlated storage. The most expensive option is decentralized storage, response by provider with relatively high costs for separate technical infrastructures and manual processing of the requests. The least expensive option is centralized storage with direct access. The investments for storage, retrieval, and management are relatively low for the following reasons: With centralized storage, only one technical infrastructure has to be set up and operated for the entire country; the requester does not have to ask each provider for data until he has found the right one because the data is stored centrally; the requester has direct access to the data store without needing an interface operated by humans.

Implementation option		Costs in [€] over five years
Decentralized storage, response by provider	Correlated	154 800 000
	Uncorrelated	157 810 000
Decentralized storage, direct access	Correlated	141 580 000
	Uncorrelated	146 100 000
Centralized storage, direct access	Correlated	133 800 000
	Uncorrelated	135 350 000
Hybrid storage, direct access	Correlated	148 320 000
	Uncorrelated	147 340 000

Table 1.3: Results quantitative assessment

1.2.4 Summary

The assessment of the various implementation options carried out in the report [58] all point towards one solution for the Dutch telecommunications market: The data of all providers is stored at a single place, the access by the requester is automated and direct to the data store, and traffic and location data are correlated and can be requested at once. According to [58], this solution is expected to generate expenses of about €133 million for the entire Dutch market for a period of five years.

The estimated costs for the eight different implementation options vary between €133 million and €157 million for a retention period of one year for the entire Dutch market. The EU directive allows a retention period of between six months and two years. Additional costs for holding the data of another year are estimated to be about €14 million for each of the options. The reduction in costs for retaining the data half a year less are estimated to be about €7 million.

Although the centralized storage with direct access has been found to be the optimal model for the aspects the report focused on, its implementation has different implications for the actors involved (requester, provider, and the intermediary third party). The authors point out that they conducted their analysis as objectively as possible but that the decision by the government may also involve other factors. The Dutch Ministry of Justice in consultation with Ministries of the Interior, Defense, and Economic Affairs has to decide what actually will be implemented by the providers.

1.3 Definition of Services

An important question when discussing the EU Data Retention directive is which data exactly has to be retained by which party involved. The exact wording used is that “providers of publicly available electronic communication services” have to collect and store data. No additional detailed

discussion of this topic is provided.

The loose usage of the term "services" allows many different interpretations. The currently probably most widely accepted interpretation is that each service provider has to retain data for exactly the services that it offers. The fact that the Internet is based on a stack of protocols building different layers allows for different interpretations of the term "service", though. If lower levels are considered as services to all other services built on top of them, then they are subject to data retention for the provider of that service, therefore generating the need to retain extremely large amounts of data.

In this thesis it is assumed that a certain service provider may *exclusively* either be an Internet Service Provider (ISP) or a mail provider. Therefore, two separate evaluations are made here: The consequences of an implementation of the EU Data Retention directive for Internet access providers or ISPs in Chapter 3 and for Internet e-mail providers further denoted as "mail providers" in Chapter 4.

2 Technical Background

This chapter explains the technical background involved in providing Internet access and Internet e-mail services. The discussion is restricted to the most important protocols, methods, and issues needed to fully understand the ideas for data retention of Internet access and Internet e-mail data introduced in Chapters 3 and 4.

2.1 OSI Model

The Open Systems Interconnection (OSI) Basic Reference Model, or OSI seven layer model as it is often called, divides the various protocols and networking hardware components into seven different layers according to their functionality (see Figure 2.1). The model was introduced in 1983 by the International Organization for Standardization (ISO)[1], a non-governmental organization, developing industrial and commercial standards. The ISO developed most of the standards for electronic communication protocols on the Internet.

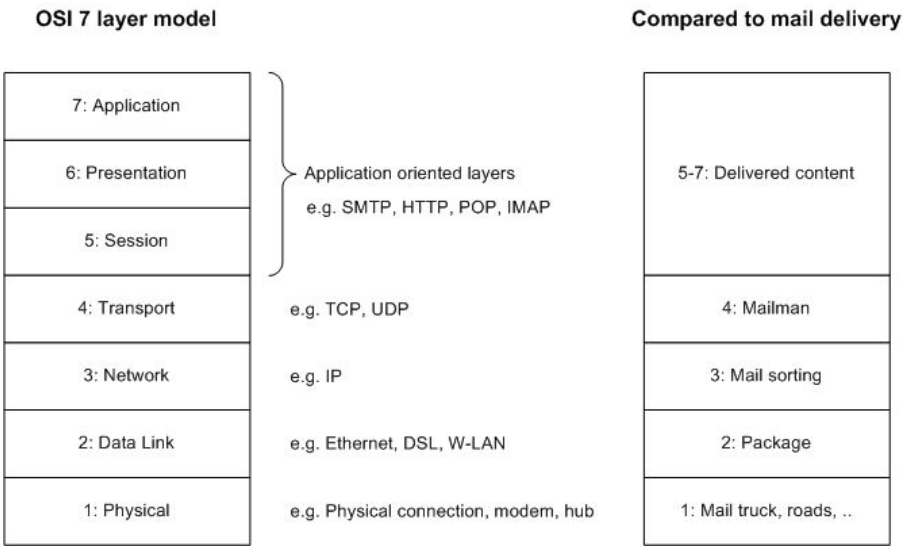

Figure 2.1: OSI 7 layer model

Every layer in the OSI model in Figure 2.1 has its responsibilities and provides its services to the layers above while relying on the layers below. Layer one, the *physical layer*, is responsible for the physical transmission

[1] http://www.iso.org

of bits without any knowledge about the information contained in these. Layer two, the *data link layer*, is the first layer to distinguish between packets of data. It is responsible for the reliable transmission of a packet stream between two physically connected devices. Devices in this context may be computers or specialized networking hardware. Layer three, the *network layer*, ensures that the packets reach their destination in an entire network. This process is called routing. Layer four, the *transport layer*, is the last layer mainly concerned with networking issues and as such provides a common interface for the application oriented layers above. It is responsible for data segmentation, flow-, and error-control. The layers five to seven, *session-*, *transportation-*, and *application*-layer are altogether often referred to as the *application oriented* layers. This is where end-user applications like e-mail clients or web browsers and their associated protocols such as SMTP, POP, and HTTP (explained in Section 2.4) reside. In general, the higher the layer, the less concerned with physical issues such as packet routing are its protocols.

The following three sections provide various examples of devices and protocols important for Internet access and Internet e-mail. The discussion is divided into three parts, according to the OSI layers involved: Layers one and two (*physical connections*), layers three and four (*TCP/IP protocol suite*), and layers four to seven (*application protocols*).

2.2 Lower OSI Layers: Physical Connections

The most popular physical Internet connections available nowadays range from dial-up modems to broadband access (see Figure 2.2). The associated devices operate on OSI layers one and two. Additionally, a short discussion of anonymous Internet access is provided in this section as this topic is of special interest in the context of the EU Data Retention directive.

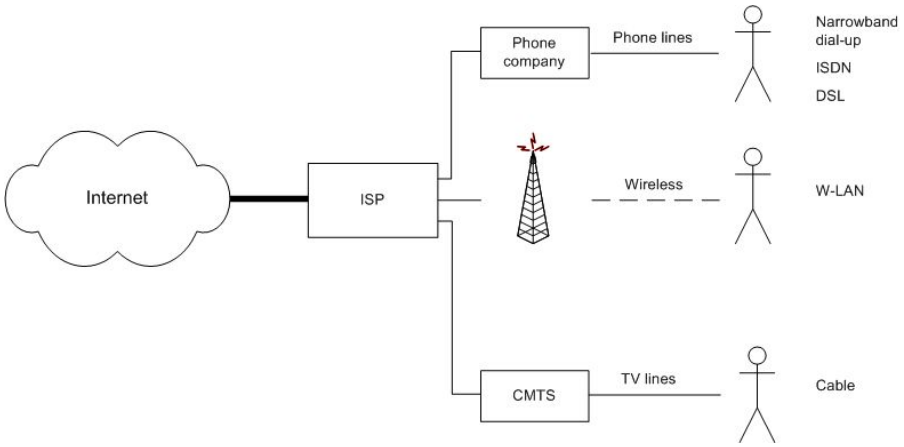

Figure 2.2: Physical connections

2.2.1 Narrowband Dial-up

Modems are devices enabling digital communication over analog Plain Old Telephone Service (POTS) lines by converting digital signals to analog and back. Narrowband modems are the predecessors of the state-of-the-art broadband Internet connections, based on the International Telecommunications Union (ITU)[2] specification V.92.

In a process called dial-up, the modem initiates the communication by trying to reach the endpoint of an Internet Service Provider (ISP) in the telephone network by dialing its number. The counterpart of the user's modem, being part of the provider's network, is connected to the Internet. Once the connection is established and the authentication was successful, a public IP address is assigned to the user's modem by the ISP-side, enabling Internet access for the customer.

2.2.2 Integrated Services Digital Network

The Integrated Services Digital Network (ISDN) provides a way to communicate digitally over preexisting telephone lines too. These copper wires were installed to be used for the landline telephone network and at that time provided a new method of digital communication without having to build a new infrastructure. In contrast to narrowband dial-up, the entire network communicates digitally and thus there is no need to convert between analog and digital signals.

2.2.3 Digital Subscriber Line

Digital Subscriber Line, or Digital Subscriber Loop (DSL) as it was originally called, is a family of standards, deployed by the European Telecommunications Standards Institute (ETSI)[3], providing yet better ways to transmit digital data over the mentioned telephone lines. By utilizing high frequency bands not being used by voice transmission it is possible to transfer digital data over these lines. It is furthermore possible to simultaneously transmit voice and data over the same line at the same time with a device called POTS splitter.

At the time of activating the DSL modem, it tries to reach its counterpart on the telephone network, the Digital Subscriber Line Access Multiplexer (DSLAM) which is connected to the ISP and represents the customer's link to the Internet. The DSLAM is, amongst other things, coordinating the overlaying of data and speech over the same line. For the user to access the Internet he has to go through some form of authentication by supplying a username and a password. Subsequently, the customer's modem is assigned a public IP address by the ISP, therefore establishing Internet connectivity.

[2] http://www.itu.int
[3] http://www.etsi.org

Asynchronous DSL (ADSL) is the most popular member of the DSL family of standards, sometimes called xDSL, and is holding 60% of the broadband marketshare worldwide [9].

2.2.4 Cable

This type of modem uses the physical lines deployed by the cable TV companies, intended to deliver television to the households. There is no need for a dial-up process, the modem at the user's home is communicating with its counterpart on the ISP side, the Cable Modem Termination System (CMTS), as soon as it is turned on. The authentication part here is done by the CMTS allowing only modems correctly configured by the provider to enter the network.

2.2.5 Wireless LAN

Since the introduction of the 802.11b standard by the Institute of Electrical and Electronics Engineers (IEEE)[4], an international non-profit organization developing specifications for communication on the Internet, which describes a method to communicate without wires, a new trend is emerging: 802.11b was followed by a suite of standards which are termed Wireless Local Area Network (W-LAN) in their entirety, improving the speed and noise reduction of wireless communication. With the hardware on the user side being built by default into most of the portable computers nowadays, the number of worldwide access points, so called "hotspots" is rising [21]. The top three hotspot locations are hotels, restaurants, and cafés [20].

The major Austrian mobile network providers are operating hotspots throughout the cities and offer mobile Internet services to their subscribed customers and to buyers of prepaid cards for time-limited access [1] [52].

Some hotspots provide their Internet access for free. This happens mostly in commercial environments, for example, the owner of a café providing free W-LAN access to his guests. These can be found using hotspot directories available on the Internet[5] [6] [7].

2.2.6 Anonymous Access

There are several circumstances under which an ISP does not possess any personal data about the customers he is serving. In these cases, what is available is some information about the network equipment. The extent to which information is available about the user's equipment depends on the type of Internet connection used.

[4] http://www.ieee.org
[5] http://www.free-hotspot.com
[6] http://www.hotspot-locations.de
[7] http://www.nodedb.com

In the situation of a customer using the free wireless Internet access (W-LAN was explained in Section 2.2.5) at a café or a restaurant for example, the ISP operating this hotspot has no personal data available on who he is connecting. What he does know is the media access control (MAC) address of the equipment being used to connect to the hotspot. A MAC address is intended to uniquely identify one specific piece of networking hardware. Each equipment vendor has its own address range which can be looked up in certain directories[8] [9]. Although the address is intended to be unique, it is forgeable using readily available software[10] [11] which renders the information available to the provider in this case unreliable.

Another possibility of anonymous Internet access is provided by certain dial-up providers (see Section 2.2.1 for details on the dial-up process). Usually, an ISP asks for a user ID and a password during the authentication process and is able to connect this user ID to an entry in his customer database therefore knowing who he is serving. In the case of the Austrian provider SelfNet[12] for example, no user-dependent credentials have to be given for authentication. All non-registered customers use the same user ID / password combination provided on the ISP's Internet site. In this situation, leaving out a traffic-analysis as information source, two things about the customer are known: The MAC address of his modem and the originating endpoint in the telephone network, the telephone number.

2.3 Middle OSI Layers: TCP/IP Protocol Suite

An inherent property of digital networks is that the transmitted packets starting from the sender pass multiple other stations before the recipient is reached. These stations may be computers or routers and are generally referred to as "hosts". The communication between these is realized by electronic protocols which standardize the way how data is carried across entire computer networks.

The protocols *Transmission Control Protocol and Internet Protocol* (TCP-/IP) have become the de-facto standards for electronic communication on the Internet in the transport and network layers of the OSI model (see Figure 2.3). The entire Internet Protocol suite was developed in the early 1970s by the Defense Advanced Research Projects Agency (DARPA)[13], an organizational party of the US defense department responsible for developing new technologies to be used by the military.

Very important for the foundation and ongoing development of the Internet are specifications formulated in Requests For Comments (RFCs).

[8] http://www.techzoom.net/mac/index.asp
[9] http://standards.ieee.org/regauth/oui/index.shtml
[10] http://www.gorlani.com/publicprj/macmakeup/macmakeup.asp
[11] http://slagheap.net/etherspoof/
[12] http://www.selfnet.at
[13] http://www.darpa.mil

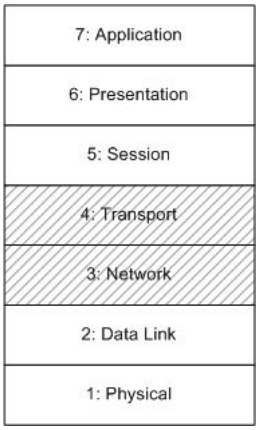

Figure 2.3: TCP/IP on the OSI 7 layer model

These documents are developed by computer experts and may be adopted as official standards by the Internet Engineering Task Force (IETF)[14], an international organization consisting of a large number of working groups for various Internet related topics. Many of the essential protocols for communication on the Internet are specified in RFCs which are freely accessible online[15].

2.3.1 Internet Protocol and Transmission Control Protocol

The Internet Protocol (IP), version six is formalized in RFC2460 [7], the older but still dominating version four is formalized in RFC791 [40], takes on the responsibility to route its data in a possibly large network from the source to a given destination host. The Transmission Control Protocol (TCP, specified in RFC793 [41]) resides on OSI layer four, the network layer, and utilizes the functions provided by the layer beneath to realize reliable data delivery. The layers three and four on the Internet are realized by IP and TCP. This protocol-bundle is often referred to as TCP/IP.

The data carried by IP is divided into packets which address a single host, whereas TCP introduces ports which provide a way to address different sites on a single host.

The general term “traffic” is used to refer to a magnitude of packets and the carried data.

2.3.2 IP Addresses and the DNS

Each IP packet contains information about its sender and the designated recipient. These IP addresses consist of 4 bytes each. Computers handle these addresses in binary representation, for human handling they are

[14] http://www.ietf.org
[15] http://rfc.net

translated to decimal numbers. For example: “192.168.0.1” is the decimal representation of the binary number “11000000 10101000 00000000 00000001”. Because these big numbers are not suited for human handling, the Domain Name System (DNS) was developed as a solution and standardized with RFC1034 [28]. The DNS standard proposes a global facility of computers responsible for translating certain domain names into IP addresses and vice-versa, so called “DNS servers” or “nameservers”. Determining the IP address associated with a domain name is called *DNS lookup* whereas the reversal process of looking up the domain name from an IP address is called *reverse DNS lookup* (rDNS). “www.google.com” is a domain name which is at the time of writing translated to the IP address “209.85.129.104”. The DNS is used every time a user enters a domain name into a browser’s address field and therefore an essential part of the current Internet.

Especially relevant in the context of this investigation are informations in the DNS which are substantial for Internet e-mail. Besides the IP address and other data, also the knowledge about the mail server responsible for the mail-handling of a certain domain is maintained by a nameserver in the so called “Mail-Exchange (MX) records”. These records are used by Mail Transfer Agents (MTAs) to acquire the IP address of the mail server responsible for a certain domain. The functionality of the global e-mail system is therefore dependent on a correctly working DNS.

2.4 Upper OSI Layers: Application Protocols

The previous two sections discussed the various networking components and protocols used to establish Internet access. This section provides various communication possibilities once a connection is established. The protocols discussed here represent only a small subset of all the protocols for electronic communication. The discussion is restricted to the most popular protocols needed to fully understand the ideas for retention of Internet e-mail data presented in Chapter 4.

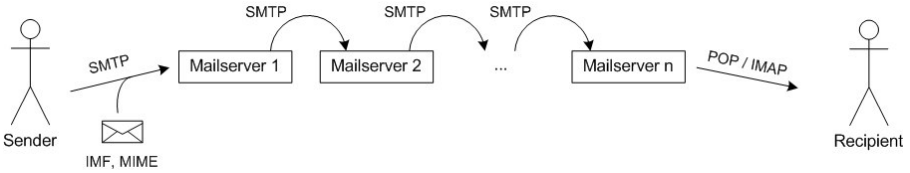

Figure 2.4: Application protocols for Internet e-mail

Figure 2.4 roughly shows the journey of an e-mail and the protocols and standards involved and discussed in the following sections: POP and IMAP for e-mail retrieval, SMTP for e-mail transmission and IMF and MIME for the formal structure of an e-mail message.

2.4.1 Simple Mail Transfer Protocol

The Simple Mail Transfer Protocol (SMTP, defined in RFC2821 [23]) is used to organize the transport of messages between two entities in the e-mail system. These may be either a mail application which is also termed Mail User Agent (MUA) or a mail server which is also termed Mail Transfer Agent (MTA).

The communication itself includes two parts: The SMTP envelope and the mail content (see Figure 2.5), which in turn consists of a header and a body. A detailed description on the Internet Message Format (IMF) is given on pp. 30.

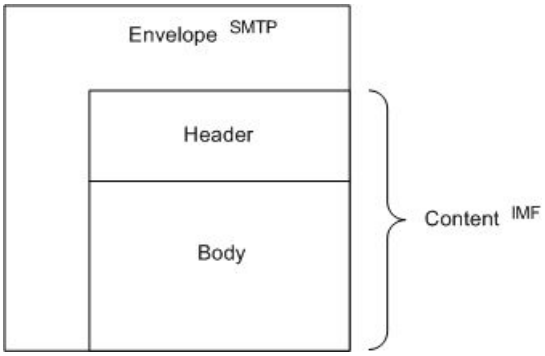

Figure 2.5: E-Mail envelope and content, see Section 2.4.4 for details

At the beginning of a mail's journey, the MUA, e.g. Microsoft Outlook[16], tries to hand it off to a mail server. It is either stored there and waits for its retrieval by the recipient or retransmitted to another mail server until the mailbox of the recipient is reached (see Figure 2.4). An electronic mailbox can be accessed via the POP or the IMAP protocol (see Sections 2.4.2 and 2.4.3 for details).

SMTP Session

In order to transmit a message via SMTP between two hosts, a two-way connection, a so called session, is established between the recipient and the sender. The two participants in a dialog are called SMTP client and SMTP server, the client being the one who has initiated the TCP connection. A session proceeds as follows: The SMTP client issues various defined commands which are read, processed, and answered by the SMTP server. The entire dialog is specified to be human-readable, the individual commands are discussed in the following section.

[16] http://office.microsoft.com/de-ch/outlook/default.aspx

Client Commands

The SMTP dialog is divided into three phases: Handshake, mail transmission, and closing of connection. See Table 2.1 for the typical SMTP commands used in each stage. Each of the commands may only be used in a certain stage.

Stage	SMTP commands
Handshake	HELO, EHLO
Mail transmission	MAIL, RCPT, DATA
Closing	QUIT

Table 2.1: SMTP commands

HELO Short for "hello", the first command issued in an SMTP session, telling the client's globally valid domain name, the so called fully qualified domain name (FQDN) of the SMTP client. Domain names are part of the Domain Name System (DNS) which is described at the end of this section.

This command is defined in RFC821 [42]. RFC2821, which introduces `EHLO` as a replacement, orders that only old software implementations may use `HELO`.

EHLO The command "extended hello", issued instead of `HELO` by the SMTP client, states the same as `HELO` but additionally informs the server that the client does support SMTP service extensions, a concept explained later in this section. In reply to this command, the server advertises his supported extensions.

Example:

```
1 Server (S): 220 foo.com SMTP Service Ready
2 Client (C): EHLO bar.com
3 S: 250 foo.com Hello bar.com [111.222.333.444]
4 S: 250 AUTH CRAM-MD5 LOGIN PLAIN
```

In this case AUTH, CRAM-MD5, LOGIN, and PLAIN are the SMTP extensions offered by this SMTP server which are not explained into further detail here.

The client has to provide its domain name or IP address along with the commands `HELO` and `EHLO`. One of the two commands has to be issued by the client as its first command in the dialog. The server may verify the domain name or IP address by performing a reverse DNS lookup for the client's IP address (see Section 2.3.1 for a description of IP addresses and DNS) obtained from the TCP connection

data. This is demonstrated in the above examples: The IP address "111.222.333.444" (line 3 in the example above) from the server's reply belongs to the client and the domain name "bar.com" (line 3) is the result of the reverse DNS lookup. This would mean that if the resolved domain name does not match the domain name given by the client (line 2), the client may be hiding its identity. Nevertheless, RFC2821 [23] rules that SMTP servers are not allowed to reject the transferred e-mail if this verification fails.

MAIL This command initiates the actual mail transaction and tells the sender's e-mail address.

Example [23]:

```
C: MAIL FROM:<Smith@bar.com>
S: 250 OK
```

The address stated by the client is the designated sender mailbox ("Smith@bar.com" in the above example), which may be chosen arbitrarily by the client. The SMTP server may choose to accept or reject the specified sender domain depending on if it is responsible for it or not, but without any SMTP extensions for security employed it is not able to verify if the SMTP client is authorized to send e-mail messages specified to originate from the given sender mailbox. The common mail transfer agents only check for syntactical correctness and therefore enable the SMTP client to provide potentially arbitrary e-mail addresses at this point.

RCPT Specifies the recipient of this message. If more than one recipient is intended, multiple `RCPT` lines have to be stated by the client. The server either replies with a 250 OK line, telling it has accepted the recipient's address, or returns a 550 reply saying that the e-mail is not deliverable to the given address.

Example [23]:

```
C: RCPT TO:<Jones@foo.com>
S: 250 OK
```

Note: although the recipient addresses are also stated in the message header of an e-mail, delivery of a message is based on the parameter of the `RCPT` command, regardless of what the message header says about the recipients. See the paragraph on processing of multiple recipients later in this section for further details.

DATA This command initiates the transmission of the actual mail content. The server has to reply with a non-error code before the client can start to send the data.

Example [23]:

```
1       C: DATA
2       S: 354 Start mail input; end with <CRLF>.<CRLF>
3       C: Date: Thu, 21 May 1998 05:33:29 -0700
4       C: From: John Q. Public <JQP@bar.com>
5       C: Subject:  The Next Meeting of the Board
6       C: To: Jones@xyz.com
7       C:
8       C: Bill:
9       C: The next meeting of the board of directors will be
10      C: on Tuesday.
11      C:  John.
12      C: .
13      S: 250 OK
```

By issuing the `DATA` command in line one of the above example, the client asks the server if he is ready to receive the message content. The server's reply in line two is positive, telling the client to start the transmission. It is immediately followed by the e-mail content seen in lines three to twelve.

The data transmitted here is the actual content of an e-mail as the recipient receives and reads it in his mail user agent or browser (web mail will be explained in Section 2.4.6) is used. It is referred to as the message content (see Figure 2.5) which consists of a header and a body which are both formatted according to certain standards explained later in the context of the Internet Message Format (IMF, see Section 2.4.4).

QUIT Tells the server that the client has no more e-mail messages to transmit and wants to end the communication.

Example [23]:

```
C: QUIT
S: 221 foo.com Service closing transmission channel
```

Message Envelope and Content

On the basis of the SMTP dialog, two central definitions are made. The envelope of an e-mail consists of the data passed by the `MAIL` and `RCPT` commands, the source and destination address respectively. The header part of the message content is intended to store sender and recipient e-mail addresses too, but regardless of the information present, only the envelope is utilized for delivery. This has important implications for blind carbon copies (BCC, described later in this section), for the Post Office Protocol (POP, see Section 2.4.2), and the Internet Message Access Protocol (IMAP, see Section 2.4.3) which are both used to access successfully delivered e-mail messages.

Although the header and the body, the two building blocks of the message content, should adhere to certain standards, RFC2821 orders that MTAs should not reject messages based on perceived defects in the message content. This therefore enables the clients of such mail servers to transmit some arbitrary header information.

Although the currently used mail user agents like Microsoft Outlook and Mozilla Thunderbird[17] do not allow the user to separately specify the value for `From` in the message envelope and header it is still not hard to introduce a customized header. A Telnet tool for example allows to open a TCP connection to a specified host and port and can be used to connect to the standard SMTP port 25 to manually manage a SMTP dialog and issue the commands by typing (which would normally have been done by the mail user agent). Telnet tools are used to open a TCP connection to a certain host and port and transmit plain-text messages via the keyboard which is very useful when debugging applications employing plain-text protocols. Introducing custom mail headers may be achieved this way but can also be done by a custom-built application.

Content Header Processing

The SMTP RFC2821 specifies that each participant acting in the role of the SMTP server in an SMTP dialog has to transcribe certain details from the dialog to the message header of a received e-mail. These details include the identity of the client, the server, and the time and date of the message transmission. Referred to as identity here are actually the Internet Protocol (IP) addresses of the client and the server whereas the former is known from the TCP connection data and the latter is the server's own address. These three pieces of information in the message header are referred to as *trace-*, *received-*, or *time stamp*-lines and are prepended to the header of the received message by the mail server. Additionally, information on the protocol, the link, the message id, and the recipients given by the `RCPT` command of the originating SMTP session may be contained. The following is one sample of a received line.

```
Received: from sr-wpay004 (sr-wpay004.smf.ebay.com
[10.10.189.14]) by mx14.sjc.ebay.com (8.13.5/8.13.5)
with ESMTP id l134mdsF27237; Fri, 2 Feb 2007 06:08:48 -0700
```

The above trace line results from a SMTP dialog between a host with the domain name "mx14.sjc.ebay.com" as the SMTP server and a host which identified himself as "sr-wpay004" during the `EHLO` command and connected with the IP address "10.10.189.14" which resolved to the domain name

[17] http://www.mozilla.com/en-US/thunderbird/

"sr-wpay004.smf.ebay.com" in the reverse DNS lookup. The SMTP session happened on Friday, February 2^{nd} of 2007 at around 6am in the local time zone (around 1pm Universal Time).

Mail servers following the RFC specification are bound to write these trace lines to the e-mail header, which were originally intended for debugging purposes, but forbidden to manipulate already existing header information. They are also not allowed to reject any e-mail based on the format of these trace fields and should be extremely robust concerning the information present in the header.

Apart from adding trace lines, there is another occasion where a mail server is obliged to add information to the message content: According to the RFC, the mail server making the final delivery has to add a *return-path line* to the beginning of the header. What is referred to as final delivery in the specification means that an e-mail message has reached the SMTP server from where it is not further transmitted via SMTP but collected by the user via POP, IMAP (see Sections 2.4.2 and 2.4.3), or similar. See the following listing for an example line.

```
Return-Path: <sender@somedomain.com>
```

The return-path line is intended to preserve the e-mail address given in the `MAIL` command of the SMTP dialog.

The resulting header of an e-mail successfully delivered to its final mail server should therefore contain exactly one return-path line and one or more received lines if all the participating electronic parties correctly processed the message according to the RFC.

The important message here for this work is that a lot of useful information is available in the header if all involved parties adhere strictly to the RFCs. It is although not possible to prove if the header information written by a foreign mail server is correct and there may even be trace lines in the header referring to SMTP dialogs that never happened which may be done to hide the real origin. Additionally, the mail servers are not allowed to reject an e-mail transmission based on the message header and therefore messages containing arbitrary header information reach their destinations.

Multiple Recipients and Carbon Copies

The processing of multiple recipients and carbon copies (CC), eventually blind carbon copies (BCC), is based on the *decoupling of header and envelope information*. The sender and recipient addresses of an e-mail message are stored in two locations: In the envelope and in the header. Only the information present in the envelope is used to deliver the message content and eventually error notifications (see commands `MAIL` and `RCPT` described above). The message header contains sender and recipient addresses too but for a different purpose: These addresses are displayed

to the user receiving the message and are not used for delivery by the mail servers.

The following SMTP dialog (partly taken from RFC2821 [23]) begins right after the client has initiated the TCP connection and shows the transmission of an e-mail message to the addresses "Jones@XYZ.COM" (line six) and "James@XYZ.COM" (line eight). The mail content transmission starts at line 12, ends at line 21, and includes the message header (lines 12 to 15) which is the interesting part in this example. The header declares "Jones@XYZ.COM" as the only recipient (line 15) of this e-mail although the content is also transmitted to "James@XYZ.COM". The mail user agent accessing the mailbox of "Jones@XYZ.COM" will declare this e-mail as being sent to "Jones@XYZ.COM" only.

```
1  S: 220 foo.com Simple Mail Transfer Service Ready
2  C: EHLO bar.com
3  S: 250-foo.com greets bar.com
4  C: MAIL FROM:<JQP@bar.com>
5  S: 250 OK
6  C: RCPT TO:<Jones@XYZ.COM>
7  S: 250 OK
8  C: RCPT TO:<James@XYZ.COM>
9  S: 250 OK
10 C: DATA
11 S: 354 Start mail input; end with <CRLF>.<CRLF>
12 C: Date: Thu, 21 May 1998 05:33:29 -0700
13 C: From: John Q. Public <JQP@bar.com>
14 C: Subject:  The Next Meeting of the Board
15 C: To: Jones@xyz.com
16 C:
17 C: Bill:
18 C: The next meeting of the board of directors will be
19 C: on Tuesday.
20 C:  John.
21 C: .
22 S: 250 OK
23 C: QUIT
24 S: 221 foo.com Service closing transmission channel
```

Extension for Authentication

The SMTP service extensions framework, sometimes referred to as Extended SMTP (ESMTP), specified in RFC1869 [24], is a collection of guidelines on how to extend the original SMTP protocol.

The SMTP AUTH extension defined in RFC2554 [32] is a way for the SMTP client to authenticate himself against the SMTP server. This authentication process can happen by asking a question to the client which

he has to answer correctly or by simply supplying a username / password pair to the server. Without the use of any other SMTP extensions, the credentials are transmitted in plain-text. This extension is a way for the server to make sure who the client is, it does, however, not provide a guarantee for the correctness of the mail header data.

Every modern e-mail client supports the SMTP AUTH extension, prompting the sender for a username / password pair if the server requires it.

See the following listing for an example SMTP dialog (partly taken from [32]) including SMTP AUTH authentication. The dialog begins right after the client has initiated the TCP connection.

```
1  Server (S): 220 smtp.example.com ESMTP server ready
2  Client (C): EHLO jgm.example.com
3  S: 250-smtp.example.com
4  S: 250 AUTH CRAM-MD5 LOGIN PLAIN
5  C: AUTH LOGIN
6  S: 334 VXNlcm5hbWU6
7  C: VGVzdFVzZXI=
8  S: 334 UGFzc3dvcmQ6
9  C: VGVzdFBhc3N3b3Jk
10 S: 235 ok
```

The actual authentication process in the listing above takes place in the lines five to ten. The lines six to nine contain encoded words in the usually human-readable SMTP dialog. These words are transformed with the Base64 character encoding specified in RFC4648 [22]. It is not a cryptographic encoding algorithm but an algorithm used to project a big alphabet onto a smaller one. Base64 is usually used to embed one protocol into another protocol or data format which both use the same vocabulary and may interfere with each other. The encoding / decoding scheme is publicly available in the specification and the transformation in both directions is accomplished without information loss. The following listing shows the mentioned lines in Base64-decoded form:

```
5  C: AUTH LOGIN
6  S: 334 Username:
7  C: TestUser
8  S: 334 Password:
9  C: TestPassword
10 S: 235 ok
```

Extension for Encryption

This service extension, defined in RFC2487 [19], is proposing a way to incorporate Transport Layer Security (TLS, see Section 2.5.2) into the

SMTP protocol. TLS provides encryption and optionally server and / or client authentication by introducing the command `STARTTLS`. If the server advertises the STARTTLS extension in his `EHLO` reply, the client may issue the identical command. Following is the TLS handshake where authentication is done if needed and the encryption is negotiated. After establishing the secured connection, the SMTP protocol starts over and the client issues the `EHLO` command again but with the entire following session being encrypted instead of a plain-text transmission.

SMTP and the DNS

The Domain Name System, defined in the RFCs 1034 and 1035 [28] [29], is one of the fundamental parts of the Internet. Its main purpose is to translate human-readable domain names like "www.google.com" to machine-readable IP addresses like "209.85.129.147" and vice-versa. Determining the IP address associated with a domain name is called DNS lookup whereas the reversal process of looking up the domain name from an IP address is called reverse lookup (rDNS). It is based on a hierarchical and distributed approach with at least two domain name servers being responsible for one particular domain. Besides the IP address and other information, also the knowledge about the mail server responsible for the mail-handling of a certain domain is maintained by a nameserver in the so called "Mail-Exchange (MX) records". These records are used by mail transfer agents to acquire the IP address of the mail server responsible for a certain domain. The functionality of the global e-mail system is therefore dependent on a correctly working DNS.

2.4.2 Post Office Protocol

The Post Office Protocol (POP), RFC1939 [33], in its current version three is designed to basically download and delete e-mail messages from a mail server. IMAP is a more advanced protocol to fulfill the same purpose and is discussed in Section 2.4.3.

Client Commands

POP3 is implemented into the currently most-used mail user agents. It is a relatively simple protocol for mailbox manipulation as can be seen in the following list of the most important commands.

STAT Asks the server for the message count in the mail-store and the used disk space.

LIST Orders the server to display a list of the stored messages.

RETR Retrieves one of the messages in the mail-store.

DELE Marks a message as trash which leads to it removal at the end of the session.

QUIT Ends the session and orders the server to delete all e-mail messages marked as trash.

Authentication is, compared to SMTP, a built-in function and has to be accomplished before a user can access his mail-store. The client commands for mailbox manipulation are answered by the server with single- or multi-line responses. See the following listing for an example POP dialog [33].

```
Server (S): +OK POP3 server ready
<1896.697170952@dbc.mtview.ca.us>
Client (C): APOP mrose c4c9334bac560ecc979e58001b3e22fb
S: +OK mrose's maildrop has 2 messages (320 octets)
C: STAT
S: +OK 2 320
C: LIST
S: +OK 2 messages (320 octets)
S: 1 120
S: 2 200
S: .
C: RETR 1
S: +OK 120 octets
C: Subject: test message
C: From: sender@mydomain.com
C: To: Y@example.com
C:
C: Hello,
C: Goodbye.
C: .
C: DELE 1
S: +OK message 1 deleted
C: DELE 2
S: +OK message 2 deleted
C: QUIT
S: +OK bye
```

Each e-mail in the server's repository for the current user is assigned a number which is used by the client to refer to it with the `RETR` and `DELE` commands.

Authentication

Before being able to access his e-mail messages, a user has to authenticate himself using the commands `USER` and `PASS` or `APOP`. The user sitting in front of the computer does not issue the commands himself, this is

actually done by the mail user agent after the latter has acquired the user's credentials with a dialog window or similar means.

Authentication via `USER` and `PASS` proceeds as follows (partly taken from [32]).

```
C: USER mrose
S: +OK mrose is a real hoopy frood
C: PASS secret
S: +OK mrose's maildrop has 2 messages (320 octets)
```

By issuing the `USER` command first the client tells the server which mailbox he wants to authenticate for (`mrose` in the above example). With the `PASS` command he transmits the plain-text password for the server to authorize the access to his mailbox.

Using the `APOP` command is an alternative authentication method which prevents the user password from being transmitted in plain-text. The client issues two parameters along with `APOP`: An identifier for the mailbox he is trying to access and a so called digest. The provided digest consists of the time stamp given by the server at the initial greeting and the password associated with the mailbox, both encrypted.

The encryption is accomplished by the use of Message-Digest Algorithm 5 (MD5), a cryptographic hash-function. Hash-functions are mathematical algorithms, also referred to as one-way functions, transforming a certain input into an output in such a way that makes it very hard to recover the input from the output. MD5 transforms every input, regardless of its size, into a 128-bit output. By applying the same MD5 function to the time stamp and the password stored on the mail server and comparing it to the digest issued by the client, the server is able to check if the user knows the password. The idea here is that a possible third party intercepting the entire dialog can not make use of the transmitted information.

See the following listing for an example POP dialog (partly taken from RFC2554 [32]) including `APOP` authentication. The dialog begins right after the client has initiated the TCP connection.

```
S: +OK POP3 server ready <1896.697170952@dbc.mtview.ca.us>
C: APOP mrose c4c9334bac560ecc979e58001b3e22fb
S: +OK maildrop has 1 message (369 octets)
```

"c4c9334bac560ecc979e58001b3e22fb" is the computed MD5 value of the input string "<1896.697170952@dbc.mtview.ca.us>tanstaaf" and "tanstaaf" is the password for the accessed mailbox.

Encryption

Transport Layer Security (TLS, see Section 2.5.2) is a popular method for applying an encryption layer to already existing protocols. The RFC2595

[35] proposes a way to apply TLS to POP and to IMAP (introduced in Section 2.4.3). A POP client can start to negotiate a TLS session by issuing the command `STLS`, short for "start TLS". This extension is supported by all popular e-mail clients like Mozilla Thunderbird or Microsoft Outlook.

SMTP Envelope

The SMTP envelope fields `MAIL` and `RCPT` of an e-mail are not transmitted to a mail client via POP. POP is designed to store the message content only, without the envelope. This implies that the only information available on sender and recipient is in the fields `From`, `To`, and `return-path` of the e-mail header. According to the SMTP RFC2821 [34] and mentioned in Section 2.4.1, header information is left unparsed and unchecked by the transmitting mail servers, can easily be modified somewhere on its way without affecting the delivery, and is therefore not to be trusted.

2.4.3 Internet Message Access Protocol

The Internet Message Access Protocol (IMAP) in its version four is defined in RFC1370 [4]. Essentially, it serves the same purpose as the POP3 protocol but in addition provides more complex organization and manipulation of the e-mail repository. Whereas POP3 is intended as a protocol for short-term storage of unread messages, IMAP is designed to access mail servers holding up to thousands of e-mail messages organized in separate folders for permanent storage.

The communication happens during a plain-text session where the client issues various commands and the server replies. The session can be secured using Transport Layer Security (TLS) for encryption and / or authentication.

Client Commands

The IMAP session is divided into four states: Not authenticated state, authenticated state, selected state, and logout state. The session proceeds from one state to the other by the client issuing state-specific commands and the server successfully processing them. The following client commands are permitted in the not authenticated state.

STARTTLS Start to negotiate a Transport Layer Security (TLS, see Section 2.5.2) session.

AUTHENTICATE Start authentication and employ data security using a Simple Authentication And Security Layer (SASL, specified in RFC2222 [31]) mechanism. SASL is a framework for designing authentication and security mechanisms based on server challenges and client replies.

LOGIN Start authentication using a plain-text username / password pair.

Commands for the authenticated state:

SELECT, EXAMINE Selects the mailbox to access. The `EXAMINE` commands does the same but opens the given mailbox read-only.

CREATE Create a mailbox.

DELETE Delete a mailbox.

RENAME Rename a mailbox.

Commands for the selected state:

CLOSE Delete all messages marked for removal and switch back to authenticated state.

SEARCH Search for a message in the current mailbox satisfying the given criteria.

FETCH Retrieve parts of or the entire data of the specified messages.

STORE This command modifies the flags for a given message or multiple messages. Certain flags like "seen" or "answered" for example can be assigned to any message.

The logout state is entered when the client issues the `LOGOUT` command or the server sends a `BYE` message which usually happens as a response to `LOGOUT`.

IMAP servers are capable of processing complex mailbox and message manipulation instructions which are in detail of limited interest for this work, therefore only the commands considered important for understanding the basic workings of this protocol have been described above.

See the following listing for an example IMAP dialog [45].

```
1  Server (S): * OK IMAP4rev1 Service Ready
2  Client (C: a001 login mrc secret
3  S: a001 OK LOGIN completed
4  C: a002 select inbox
5  S: * 18 EXISTS
6  S: * FLAGS (\Answered \Flagged \Deleted \Seen \Draft)
7  S: * 2 RECENT
8  S: * OK [UNSEEN 17] Message 17 is the first unseen message
9  S: * OK [UIDVALIDITY 3857529045] UIDs valid
10 S: a002 OK [READ-WRITE] SELECT completed
11 C: a003 fetch 12 full
12 S: * 12 FETCH (FLAGS (\Seen) INTERNALDATE
13 "17-Jul-1996 02:44:25 -0700"
```

```
14      RFC822.SIZE 4286 ENVELOPE
15       ("Wed, 17 Jul 1996 02:23:25 -0700 (PDT)"
16      "IMAP4rev1 WG mtg summary and minutes"
17      (("Terry Gray" NIL "gray" "cac.washington.edu"))
18      (("Terry Gray" NIL "gray" "cac.washington.edu"))
19      (("Terry Gray" NIL "gray" "cac.washington.edu"))
20      ((NIL NIL "imap" "cac.washington.edu"))
21      ((NIL NIL "minutes" "CNRI.Reston.VA.US")
22      ("John Klensin" NIL "KLENSIN" "MIT.EDU")) NIL NIL
23      "<B27397-0100000@cac.washington.edu>")
24       BODY ("TEXT" "PLAIN" ("CHARSET" "US-ASCII") NIL NIL
25       "7BIT" 3028 92))
26  S: a003 OK FETCH completed
27  C: a004 fetch 12 body[header]
28  S: * 12 FETCH (BODY[HEADER] {342}
29  S: Date: Wed, 17 Jul 1996 02:23:25 -0700 (PDT)
30  S: From: Terry Gray <gray@cac.washington.edu>
31  S: Subject: IMAP4rev1 WG mtg summary and minutes
32  S: To: imap@cac.washington.edu
33  S: cc: minutes@CNRI.Reston.VA.US, J. Klensin <KLENSIN@MIT.EDU>
34  S: Message-Id: <B27397-0100000@cac.washington.edu>
35  S: MIME-Version: 1.0
36  S: Content-Type: TEXT/PLAIN; CHARSET=US-ASCII
37  S:
38  S: )
39  S: a004 OK FETCH completed
40  C: a005 store 12 +flags \deleted
41  S: * 12 FETCH (FLAGS (\Seen \Deleted))
42  S: a005 OK +FLAGS completed
43  C: a006 logout
44  S: * BYE IMAP4rev1 server terminating connection
45  S: a006 OK LOGOUT completed
```

The above sample dialog is composed as follows. Lines $1-3$: Not authenticated state; lines $4-10$: Authenticated state; lines $11-40$: Selected state; lines $41-43$: Logout state.

Encryption

Transport Layer Security (TLS, see Section 2.5.2) is a popular method for applying an encryption layer to already existing protocols. The RFC2595 [35] proposes a way to apply TLS to POP and IMAP. An IMAP client can start to negotiate a TLS session by issuing the command `STARTTLS`, introduced in RFC2595. This extension is supported by all popular e-mail clients like Mozilla Thunderbird or Microsoft Outlook.

SMTP Envelope

The SMTP envelope fields `MAIL` and `RCPT` of an e-mail are not transmitted to a mail client via IMAP. IMAP is designed to store the message content only, without the envelope. This implies that the only information available on sender and recipient is in the fields `From`, `To`, and `return-path` of the e-mail header. According to the SMTP RFC2821 [34] and mentioned in Section 2.4.1, header information is left unparsed and unchecked by the transmitting mail servers, can easily be modified somewhere on its way without affecting the delivery, and is therefore not to be trusted.

2.4.4 Internet Message Format

The format of an e-mail itself as it is transported by the protocols above has to fit the specification of the Internet Message Format (IMF), defined 2001 in RFC2822 [45]. Each message consists of an envelope and a content. The envelope is defined by the SMTP protocol. The content, which is subdivided into header and body, is defined by the IMF (see Figure 2.5 in Section 2.4.1).

Header

The header is generally made up of unordered name / value-pairs which are divided into several different groups according to their semantics. The idea is to provide the recipient with various additional details apart from the content. There is no restricted list of possible header fields, additional header fields may be defined [39] and sent along without violating the RFC. See Table 2.2 for a complete list of compulsory and optional header fields defined in RFC2822.

The following is an example of a message content which is comprised of header and body [45].

```
1  From: Mary Smith <mary@example.net>
2  To: John Doe <jdoe@machine.example>
3  Reply-To: "Mary Smith: Personal Account" <smith@home.example>
4  Subject: Re: Saying Hello
5  Date: Fri, 21 Nov 1997 10:01:10 -0600
6  Message-ID: <3456@example.net>
7  In-Reply-To: <1234@local.machine.example>
8  References: <1234@local.machine.example>
9
10 This is a reply to your hello.
```

The header in the example message above is composed of lines $1 - 8$ and the body of the remaining lines.

The fields `Date` and `From` are compulsory for a header to be valid, the others are optional. The `Date` field has to contain the date and time when

Group	Fields
Originator fields	From, Sender, Reply-To
Destination address fields	To, Cc, Bcc
Identification fields	Message-ID, In-Reply-To, References
Informational fields	Subject, Comments, Keywords
Resent fields	Resent-Date, Resent-From, Resent-Sender, Resent-To Resent-Cc, Resent-Bcc, Resent-Message-ID
Trace fields	Return-Path, Received

Table 2.2: Header fields

the message was ready and ordered to be sent by the user. `From` contains one or more mailboxes giving the author's e-mail addresses and their real names which may be chosen arbitrarily.

There is a huge number of possible header fields nowadays for an e-mail message. On the one hand, a number of standards define new fields and on the other hand, a lot of popular applications introduce non-standard fields. The latter is possible because the RFCs do not prescribe the standardization of all headers being used, RFC822 [6] only ordered this user-defined fields to start with "X-". Although RFC822 became obsolete with RFC2822 which did not mention the prefix "X-", it is still common to use it. There are efforts to provide an overview of the fields used like a central registry with RFC3864 [25] or a composition of the most common ones[18].

Body

The body contains contains the actual message the user typed in his mail user agent, it can be enriched in a way described in the five Multipurpose Internet Mail Extensions (MIME) RFCs 2045 [15], 2046 [16], 2047 [30], 2048 [17], and 2049 [14].

SMTP Envelope

It is important to note again that the header fields `To`, `Cc`, and `Bcc` are *not used for delivery* by the mail transfer agents. In the same fashion, the `From`, `Sender`, and `Reply-To` fields are not used for unsuccessful delivery notifications by an MTA. The header fields are used for being displayed to the user only whereas the SMTP envelope is used to deliver the message.

[18] http://people.dsv.su.se/ jpalme/ietf/mail-headers

2.4.5 Hypertext Transfer Protocol

The Hypertext Transfer Protocol (HTTP), which is defined in RFC2616 [12] in its version 1.1, is used to transfer web pages across the Internet. Web pages are written according to the Hypertext Markup Language (HTML [44]), optionally with multimedia embedded. The HTML standard was initially developed and is continuously extended by the World Wide Web Consortium (W3C)[19], an international organization with the mission to develop and improve the protocols and formats for communication on the Internet. Visiting a website basically basically means viewing the browser's interpretation of a HTML page which got transferred from a web server using the HTTP protocol.

A HTTP session starts with the client opening a TCP connection to the server and sending one or more requests to the server. The request commands as well as the reply codes are defined in the RFC. The currently most-used commands are the two following.

GET The usual request method, simply asking for a resource on the server.

POST Same as get, but allows to transfer name / value pairs as parameters within the requests.

The other client commands defined in the RFC are `PUT`, `HEAD`, `DELETE`, `TRACE`, `OPTIONS`, and `CONNECT`. They are not mentioned or explained further in this section as they are left unused by the current web browsers and are not important to show the basic workings of HTTP.

The following listing is a sample client request.

```
1  GET / HTTP/1.1
2  Host: www.google.at
3  User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US;
4  rv:1.8.1.3; Google-TR-3) Gecko/2007
5  Accept: text/xml,application/xml,text/html;
6  q=0.9,text/plain;q=0.8,image/p
7  Accept-Language: en-us,en;q=0.5
8  Accept-Encoding: gzip,deflate
9  Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
10 Keep-Alive: 300
11 Connection: keep-alive
12 Cache-Control: max-age=0\
```

Line one in the above sample transmission defines the method used for this request, the requested resource ("/" in this case), and the HTTP protocol version ("1.1" in this case). The remaining lines provide, amongst other things, information on the web browser used, the configured language on

[19] http://www.w3.org

the client user interface side and the accepted data formats which the browser is capable of displaying.

The following listing is a sample server response to the request above.

```
1  HTTP/1.1 200 OK
2  Request Version: HTTP/1.1
3  Response Code: 200
4  X-TR: 1
5  Cache-Control: private
6  Content-Type: text/html; charset=UTF-8
7  Content-Encoding: gzip
8  Server: GWS/2.1
9  Content-Length: 1708
10 Date: Fri, 04 May 2007 14:13:30 GMT
11 Content-encoded entity body (gzip): 1708 bytes -> 3647 bytes
12 Line-based text data: text/html
13 <html>
14 <head>
15   <meta http-equiv="content-type" content="text/html;
16    charset=UTF-8">
17 <title>
18 Google
19 </title>
20 ..
21 </head>
22 <body>
23 ..
24 </body>
25 </html>
```

Compared to the request which consists of a header solely, the response also contains a body. The header is comprised of the lines $1-12$ and gives the client information about the data format of the response, the sent data length etc. The remaining lines are occupied by the response body which is an abbreviated HTML representation of the Austrian Google front-page[20]. The message body is read by the web browser, interpreted and visually presented to the user.

2.4.6 Web Mail

Web mail providers offer their users the possibility to access and manage their mailboxes via a browser instead of a traditional mail application like Mozilla Firefox[21] or Microsoft Outlook. A web mail interface has become a natural thing for the big mail service providers in the last few years. See

[20] http://www.google.at
[21] http://www.mozilla.com/en-US/firefox/

Figure 2.6 for an overview of the process involved. See Figure 2.4 for a comparison with traditional e-mail.

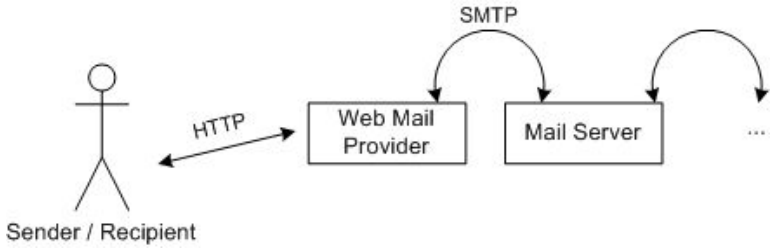

Figure 2.6: web mail providers overview

What is important in the context of this work is the use of an additional protocol with web mail. Here the user types his e-mail messages into a text box on a website and thereby sends it to the web server of the web mail provider. The web server is either a mail server too or somehow transports the message to a mail server from where it leaves on its way to the destination mailbox. When receiving a message, the process works in the opposite direction: The web mail server receives the messages and the user reads them via browser. In both cases, the communication between web mail provider and user happens via HTTP (see Section 2.4.5) compared to SMTP and POP/IMAP.

2.5 Related Issues in Electronic Communications

By sending sensitive data over multiple computers there is always the risk that these packets are observed by third parties which is called *interception*. This naturally leads to the necessity of data *encryption*. Also an interesting issue is the identity of the source of a communication and how its *anonymity* may be established.

2.5.1 Interception

As mentioned before, IP packets typically have to traverse a number of hosts on the Internet in order to reach their final destination. The process of guiding packets through a network is called "routing" and is done transparently with the user browsing the Internet not noticing that the data he requested is actually not arriving directly from the web server. The Internet Protocol (IP) is designed to decide which route to take for a certain packet in order to reach its destination. The following listing is taken from the truncated output of a traceroute application[22], showing the IP addresses of the hosts passed by a packet on its way to "www.google.com".

```
1    213.33.98.18 213.33.98.18
```

[22] http://kmu.telekom.at/kundenbereich/Internettools/Traceroute.php

```
2   80.120.165.2 80.120.165.2
3   195.3.70.29 195.3.70.29
4   195.3.70.86 195.3.70.86
5   de-cix10.net.google.com  80.81.192.108
6   209.85.249.180 209.85.249.180
7   209.85.248.248 209.85.248.248
8   72.14.239.46 72.14.239.46
9   72.14.239.58 72.14.239.58
10  mu-in-f103.google.com  209.85.135.103
```

As can be seen above, nine hosts have to be passed for the packet to reach its final destination. Specialized software or hardware components called *network analyzers* or *sniffers* are used to observe the IP packets routed through a host. These components may either be installed at a router itself or strategically placed between two hosts in order to monitor certain traffic routes.

2.5.2 Encryption

The need to encrypt certain types of electronic data naturally arises from the reasons mentioned in the previous section. In some technical setups it is especially easy to intercept network traffic. For example, when thinking of wireless communications where an intercepting party does not even need physical access to the transmission lines. Typical uses of encryption include Internet money transfers and sensitive communication in general. This section introduces a few popular algorithms and their applications in networking protocols.

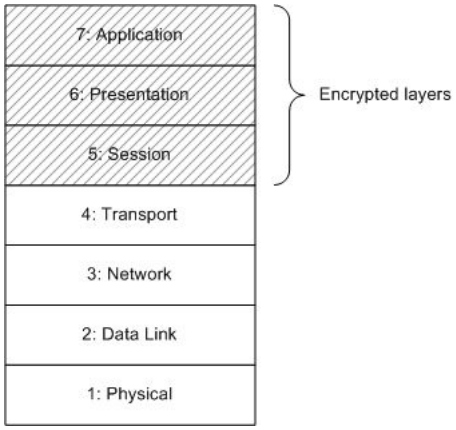

Figure 2.7: OSI layers relevant for encryption

The protocols examined for their encryption-capabilities in this section operate on the OSI layers five to seven (see Figure 2.7). The TCP/IP protocols are implementations of the layers three and five. Protocols on lower layers carry the content of higher level protocols, for example: TCP

packets (layer four) carrying encrypted application level protocols (layers five to seven) display their headers, which provides information about sender and recipient, in plain-text. A third party watching such encrypted packets is therefore able to see where the packets are coming from and where they are going to but does not know what the communication is about, given that he is not able to decrypt the content.

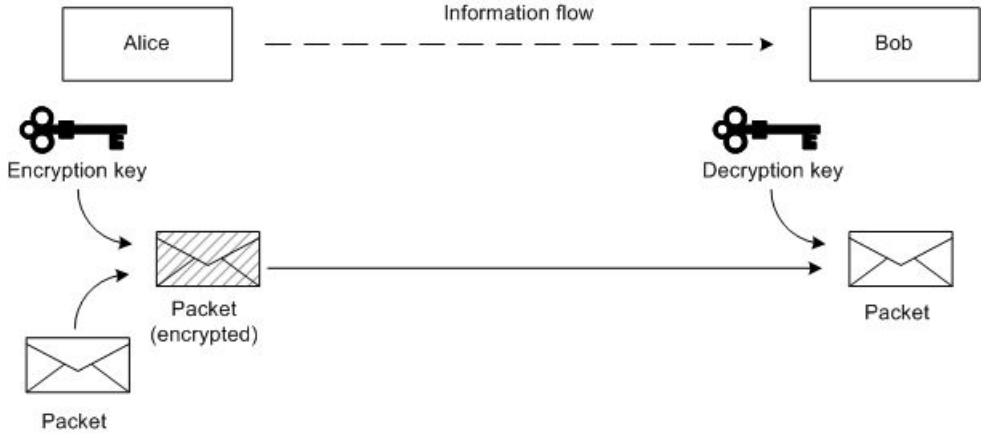

Figure 2.8: Encrypted communication scenario

Figure 2.8 shows an encrypted communication scenario between the two parties Alice and Bob. On both sides a key is needed for en- and decryption. Depending on whether the keys used are the same or different, the various algorithms are classified in two categories.

Algorithms

Symmetric: The same key is used by both parties for en- and de-cryption.

Asymmetric: Different keys are used by the two parties, which is also called public-private-key encryption.

Symmetric-key methods are often faster than asymmetric, but pose the problem of negotiating over a shared key with your communication partner without a third party knowing. An often applied technique therefore is to use a slower asymmetric-key encryption method to negotiate a shared key and then proceed symmetrically with the communication using this shared key.

A few popular members of the two categories will now be introduced shortly.

Symmetric algorithms

DES (1976) and 3DES (1978): The Data Encryption Standard (DES) was announced by the National Institute of Standards and Tech-

nology (NIST)[23] as a federal information processing standard of the United States (US) in 1976 [57]. It was revised a few times [54], and considered insecure by crypto-analysts [8] [10]. 3DES is the application of the DES algorithm for three successive times using three different keys [56]. It was developed to quickly provide an alternative after DES was declared insecure, until an appropriate successor was designed: AES.

RC4 (1987): Rivest Cipher 4 (RC4) or "Ron's Code" (after the designer Ron Rivest) [46]. The algorithm is incorporated into many software products, although some popular applications are considered to be insecure implementations [13] [59].

AES (2001): The Advanced Encryption Standard (AES) was announced by the NIST as a federal information processing standard for the US in 2001 [55].

Asymmetric algorithms

RSA (1978): Was described in 1978 and named after its developers Rivest, Shamir, and Adleman (RSA) [47]. It is used for encryption and digital signatures and considered secure if it is implemented and used correctly [5].

DSA (1994): The Digital Signature Algorithm (DSA) was announced by the NIST as a federal information processing standard of the US in 1994. It was published as part of the Digital Signature Standard (DSS) [53].

Applications

Secure Sockets Layer (SSL) and its successor Transport Layer Security (TLS) are two methods used for securing communications between a browser and a web server. TLS is intended as an application independent layer of security embedded between Transmission Control Protocol (TCP) and the actually used protocol which is the Hypertext Transfer Protocol (HTTP, described in Section 2.4.5) in the case of a secured web browser to web server communication. The operation of TLS is divided into two phases: During the handshake phase authentication is provided by using public-private key encryption and negotiating a shared secret one-session key for symmetric encryption. In the second phase, symmetric encryption is used to transport the actual application protocol data. Possible protocols used for the asymmetric part are, amongst others, RSA and DSA and for the symmetric part RC4, DES, 3DES, and AES.

Pretty Good Privacy (PGP) is a set of applications providing means for encrypted communication and digital signatures, developed by the PGP

[23] http://www.nist.gov

Corporation[24]. Again, public-private-key encryption is used to agree over a shared symmetric key which is used for further communication.

2.5.3 Anonymization

Anonymization services are companies, organizations, or distributed software tools which serve the purpose of protecting the source of a communication. This is done by adding one or more virtual layers between the source and the destination of an IP communication. This technique makes the request appear to be coming from the anonymization service instead of the real origin. These applications operate on the transport and network layers of the OSI model (see Figure 2.9).

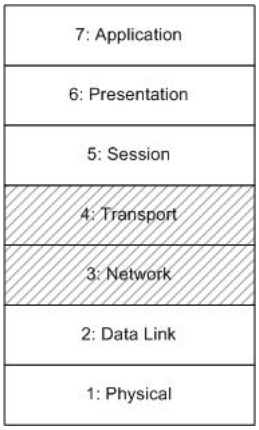

Figure 2.9: OSI layers affected by anonymization services

Within the following three sections an overview of the concepts and actual implementations of methods to anonymize web browsing, e-mail messages, and Internet traffic as a whole is provided. Although there are other technical approaches to accomplish anonymization, the ones discussed here are regarded the most important in the context of this work.

Anonymization Proxies

Anonymization proxies, in contrast to traditional proxies, do not serve the purpose of caching and serving popular files. These services ask you for the website you want to visit, load it on their behalf, and serve it back to you. The communication therefore seems to happen only between the web server and the anonymization service from the perspective of the involved web server. Example anonymization services are anonymizer[25] and Anonymization.Net[26]. See Figure 2.10 for a schematic overview of anonymization proxies.

[24] http://www.pgp.com
[25] http://www.anonymizer.com
[26] http://www.anonymization.net

Figure 2.10: Communication flow

This technique is designed to hide web traffic only. See Section 2.4.5 for a detailed description of web traffic and the Hypertext Transfer Protocol (HTTP).

Remailers

Remailers try to disguise the original sender of an e-mail message. This is done by accepting an e-mail, removing all the information about the sender, and forwarding it to the given recipient address. Different configurations exist with specialized remailers which enable bi-directional e-mail communication but still sustain the anonymity of the sender.

Anonymization Networks

The Java Anon Proxy (JAP)[27] and The Onion Router (TOR)[28] projects develop applications for so called "anonymization networks". The idea of both projects is to route Internet Protocol (IP) packets through a variety of specialized computers in order disguise the real origin of the packet. The packets are made look like they come from one of the routers of the anonymization network (see Figure 2.11). The computers responsible for routing the traffic are operating special routing software provided on the JAP and TOR project websites.

[27] http://anon.inf.tu-dresden.de
[28] http://tor.eff.org

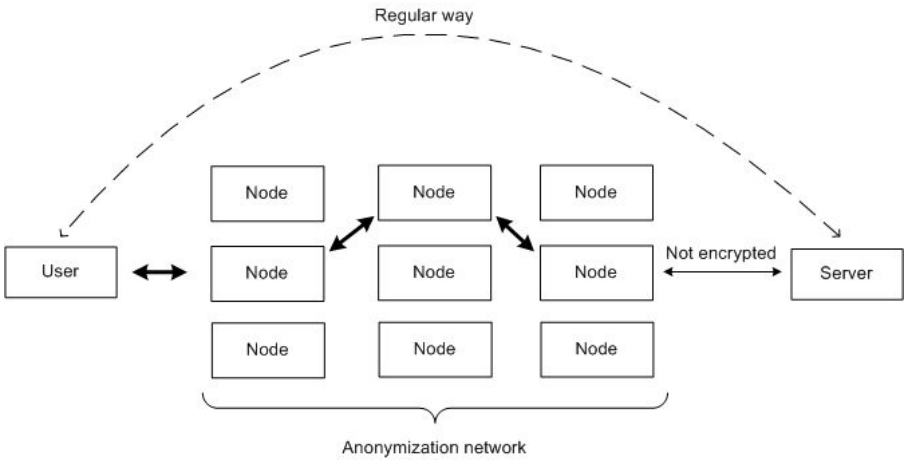

Figure 2.11: Anonymization networks

3 Data Retention for Internet Access

Internet Service Providers (ISPs) are the entry points to the global web for end-users. Although everybody is accessing the same network, there is a variety of technical ways offered by the ISPs to enable it. The types of Internet connections vary by pricing, bandwidth, and their technical setups which will be the topic of this chapter.

3.1 Requirements of the EU Directive

For Internet access, the EU Data Retention directive 2006/24/EC requires storage of the following information.

Art 5 (1) lit a Z 2 i: "data necessary to trace and identify the source of a communication: ... the user ID(s) allocated"

Art 5 (1) lit a Z 2 ii: "data necessary to trace and identify the source of a communication: ... the user ID and telephone number allocated to any communication entering the public telephone network"

Art 5 (1) lit a Z 2 iii: "data necessary to trace and identify the source of a communication: ... the name and address of the subscriber or registered user to whom an Internet Protocol (IP) address, user ID or telephone number was allocated at the time of the communication"

Art 5 (1) lit c Z 2 i: "data necessary to identify the date, time and duration of a communication: ... the date and time of the log-in and log-off of the Internet access service, based on a certain time zone, together with the IP address, whether dynamic or static, allocated by the Internet access service provider to a communication, and the user ID of the subscriber or registered user"

Art 5 (1) lit e Z 3 i: "data necessary to identify users' communication equipment or what purports to be their equipment: ... the calling telephone number for dial-up access"

Art 5 (1) lit e Z 3 ii: "data necessary to identify users' communication equipment or what purports to be their equipment: ... the digital subscriber line (DSL) or other end point of the originator of the communication"

3.2 Possible Implementation of the EU Directive

The investigation in this thesis focuses on the implications of the EU Data Retention directive including the financial costs and the additional disk space needed. In order to provide an in-depth analysis, a bottom-up approach is chosen: This sections starts by introducing a model ISP and estimating its network traffic. Furthermore, a data model design is provided and extrapolated to the whole mail provider. Finally, an estimation of the monetary costs caused by the storage and supply of the retained data follows.

3.2.1 Model ISP

In order to quantify the amount of disk space and the costs associated with satisfying the requirements of the EU Data Retention directive for Internet access a fictitious model provider with a certain set of properties is defined.

Data retention period	6 months
Request volume per year	200
Dialup customers	184 000
Broadband customers	316 000
Dialup logins per day and customer	2
Broadband logins per day and customer	1

Table 3.1: Model provider figures for Internet access

The model provider introduced here serves 500 000 customers and represents a medium-sized Austrian ISP. By extrapolating the figures in [26], the request volume for Internet access records by government authorities was assumed to be 200 per year. According to [50], 63.2% of the Austrian households are connected via broadband to the Internet and the remainder via dial-up. For an ISP serving 500 000 customers this results in 316 000 broadband accounts. Table 3.1 summarizes the figures needed for estimating the disk space.

The term "login" in Table 3.1 refers to the process of establishing a connection to the ISP. Various technical tasks have to be carried out to do this, depending on the technical equipment used by the customer. For example, a certain number has to be dialed by the modem in the case of connecting over the public telephone network. Dial-up and broadband connections differ in the speed of access provided to the customer and the underlying technical infrastructure. No official data was available on the number of logins. Therefore, two dial-up logins and one broadband login per day and customer were assumed for the model provider. See

Section 2.2 for a complete description of the most popular Internet access methods.

3.2.2 Storage Requirements

This section is concerned with the costs in terms of needed disk space. First, a data model design is presented and discussed which is used to derive a record size for the Internet access data store. Based on this figure, the needed disk space for the model ISP introduced in Section 3.2.1 is calculated.

Data Model

In order to provide a meaningful estimation of the disk space requirements a detailed data model has been worked out. The proposed data model is shown in Table 3.2.

Field	Type	Bytes needed
ID	Numeric	4
Customer_Ref	Numeric	4
Connection	Numeric	1
Log-in	Date	7
Log-off	Date	7
IP	Numeric	4
Source_Telephone_Number	Text	15

Table 3.2: Storage data model for Internet access

For each of the fields shown in Table 3.2 a disk space requirement in bytes is provided. These values are valid for the Oracle database management system[1] (see [37]).

Following is a discussion of each of the fields in Table 3.2.

ID This attribute is the so called primary key of the table which uniquely identifies exactly one record. It is not possible for two records to have the same `ID`.

Customer_Ref The existence of an ISP-internal customer database is assumed, hence, it is not necessary to store the name and address again here. Instead, only a reference, a customer ID or something similar uniquely identifying one customer, is used.

The disk space of four bytes allows for this field to theoretically hold a few billions of different integer values which seems more than enough for a customer ID field. Providers may also employ combinations of integer and character values which take up more disk space, therefore

[1] http://www.oracle.com/database/index.html

a number of bytes larger than what would be needed in theory was assumed.

Connection Because this data model is used for all the different methods of Internet access a provider serves to his customers, this attribute tells whether the connection was established through dial-up, DSL, cable modem, or something else. The idea is to assign and store numbers for each type of connection, for example "0" for dial-up, "1" for DSL, etc.

Log-in, Log-off These attributes represent the date, time, and timezone of the points in time the user established and released his connection to the Internet. Both fields are time stamps containing the date, hour, minute, and second of an event. Seven bytes are needed for each value.

IP The public Internet Protocol (IP) address (see Section 2.3.1) assigned to the user's hardware. Four bytes are needed to store it.

Source_Telephone_Number The EU Data Retention directive requires the "the calling telephone number for dial-up access" and "the digital subscriber line (DSL) or other end point of the originator of the communication" which are met by this field. According to the Austrian numbering plan, the maximum telephone number length including the country code is 15 digits [48], not taking phone extensions into account. Therefore the same amount of bytes is estimated to be needed for one record because one character takes up one byte in the data store.

A differentiation between broadband and dial-up customers has to be made for the estimation of the average record size. The field `Source_Telephone_Number` is obviously needed only for Internet access records representing dial-up connections over the public telephone network like narrowband dial-up or ISDN (see Section 2.2). Therefore, the average record size is 27 bytes for a broadband connection record and 42 bytes for a dial-up connection record.

Disk Space

The EU Data Retention directive requires the ISPs to store the data for at least six months and for at most two years. Based on the numbers provided for the model ISP in Section 3.2.1 and this section, it was estimated that an Austrian ISP serving 500 000 customers with average behavior faces a permanent additional disk space requirement of about 8.9 gigabytes for storing Internet access related data required by the EU directive for a six months period, including a full backup of the stored data.

After having evaluated the needed disk space, the next topic will be the monetary costs caused by an implementation of the EU Data Retention directive.

3.2.3 Costs

In order to provide a standardized view on this topic, the costs described apply to the model ISP introduced in Section 3.2.1.

The entire task can be subdivided into *data storage* and *data retrieval*. The *data storage* process includes the gathering of traffic and location data, possible transformations applied to the data, and the archiving of the data. *Data retrieval* refers to the information flow from the ISPs to government authorities as a reaction to specific queries.

Whereas the data storage for Internet access related data differs quite much compared to Internet e-mail data, the data retrieval does not. Data retrieval costs are roughly caused by the setup and maintenance of a server and the development of an appropriate user interface which are basically the same for the Internet access and e-mail. Therefore, the data retrieval costs are summarized here and explained in detail in Appendix B.

In terms of personnel costs, the salary charged for one technician was estimated as €120 per hour which corresponds to €19 200 per month per full time equivalent (FTE).

The numbers obtained from [51] and [38] have been rounded up in the last digit.

Data Storage Costs

The storage process includes *gathering*, *processing*, and *archiving* of the required data.

The costs were divided into setup costs incurring once (see Table 3.4) and operational costs incurring every month (see Table 3.3).

Category	Item	Costs [€/month]	Remarks
HW	maintenance of storage and acquisition server	60	e.g. Sun Gold Support (product number W9D-B15W-3G [51])

Table 3.3: Operational costs for *data storage* of Internet access related data

Category	Item	Costs [€]	Remarks
HW	storage and acquisition server	11 120	e.g. Sun Fire X4450 server (product numbers B15-VZ4-CB-8GB-JL6, X311L, SGXPCIESAS-R-INT-Z, X6388A, XRA-SS2CF-73G10K [51])
SW	database software	11 160	e.g. Oracle Database Enterprise Edition for one processor [38]
DEV	project setup, software development & deployment	288 000	15 FTEs (2.5 FTEs for 6 months)

Table 3.4: Setup costs for *data storage* of Internet access related data

Data Retrieval Costs

It is not enough to store the data to be retained. Provisions need to be made for efficiently accessing the data stored in order to respond to queries. The retrieval process includes *extraction* of the requested data from the data warehouse and *delivering* it to government authorities.

As mentioned before, the data retrieval costs are basically the same for Internet access and Internet e-mail and are therefore summarized in Table 3.5 and explained in detail in Appendix B.

	HW	SW	DEV	GEN	Σ
Setup [€]	15 990	0	115 200	0	131 190
Operation [€/year]	1 920	0	115 200	115 200	232 320

Table 3.5: Data retrieval costs (for more details, see Appendix B)

Total Costs

The data in Table 3.6 shows that the EU Data Retention directive causes overall costs of about €673 790 in the first year, and of about €232 320 in each of the following years for the model ISP introduced in Section 3.2.1.

3.2.4 Open Issues

The previous sections of this chapter provided suggestions on how the regulations of the EU directive for the area “Internet access” can be im-

	HW	SW	DEV	GEN	Σ
Setup [€]	27 110	11 160	403 200	0	441 470
Operation [€/year]	1 920	0	115 200	115 200	232 320

Table 3.6: Total costs for storage and retrieval for Internet access

plemented. This section provides an overview of the open issues that remain.

Requirements of the EU Directive

Ambiguous formulations and requirements are discussed in the following paragraphs and, additionally, assumptions are made about what could be meant by the authors of the EU directive.

Art 5 (1) lit a Z 2 i: "data necessary to trace and identify the source of a communication: ... the user ID(s) allocated"

The term "user ID" typically usually refers to a unique identifier of a certain user. In the case of no user ID being assigned by the provider, the IP address of a customer and the corresponding time it was assigned to the former does, in combination, uniquely identify a user in the network of an ISP. These two fields are marked for storage in the proposed data model in Section 3.2.2 which is therefore fulfilling the EU directive's requirement for a user ID. If there are certain user IDs assigned to each customer, for example the user name used for authorization when connecting via DSL, they must be somehow logically connected to `Customer_Ref` in the provider's internal customer database and can therefore be figured out on the basis of a data storage according to Table 3.2.

Art 5 (1) lit a Z 2 ii: "data necessary to trace and identify the source of a communication: ... the user ID and telephone number allocated to any communication entering the public telephone network"

The purpose of a user ID and its interpretation as a requirement for Internet access are discussed in the paragraph above.

Only in the case of dial-up access a telephone number is assigned, when thinking of cable modem access for example no telephone numbers are involved in the process of gaining Internet access. See Section 2.2 for the different types of Internet connections available. As all the requirements below Art 5 (1) lit a Z 2 of the EU directive are in the context of Internet access, Internet e-mail, and Internet telephony, this requirement most likely was thought to apply to Internet telephony, where the term user ID and the process of an allocated telephone number entering the public telephone network makes most sense.

Art 5 (1) lit c Z 2 ii: "data necessary to identify the date, time and duration of a communication: ... the date and time of the log-in and log-off of the Internet e-mail service or Internet telephony service, based on a certain time zone"

This requirement is formulated as a sub item of Art 5 (1) lit c Z 2 putting it in the context of Internet access, Internet e-mail, and Internet telephony. It requires information about an Internet e-mail or telephone service to be stored, but with Internet access being the focus of this section this requirement is again thought to be applying to Internet e-mail and telephony and is therefore not further considered in the context of Internet access.

This requirement is being considered in Chapter 4 for Internet e-mail.

Anonymous Access

Considering the case of somebody using a wireless hotspot with the prepaid cards mentioned in Section 2.2.5: The ISP which is connecting the hotspot has no data about this customer except his MAC address (see Section 2.2.6).

Another case of anonymous Internet access is provided by ISPs offering dial-up access without previous registration. As with SelfNet [49] the user is using a dial-up number, login, and password provided at a publicly accessible homepage.

3.3 Conclusion

The fraction of broadband in overall Internet connections in Austria is rising [50] and the fraction of people using dial-up access is falling. Dial-up connections usually are billed according to the time connected to the ISP, broadband connections by the amount of data transferred. When being billed according to data volume there is no need for the customers anymore to disconnect after use, so online time, which is required to be retained according to the EU Data Retention directive, does not imply online activity.

The overall storage requirement caused by the EU Data Retention directive in the area of Internet access for the model ISP with 500 000 customers defined in Section 3.2.1 is estimated to be around 8.9 gigabytes (including backup). The monetary costs caused by an implementation are estimated to add up to €673 790 in the first year, and €232 320 in the following years.

Depending on the types of Internet connections provided by an ISP, the data required by the EU directive to be stored concerning Internet access is to a large extent available. There are although easily accessible ways for people with basic technical knowledge to gain Internet access and stay anonymous.

4 Data Retention for Internet E-Mail

The global e-mail system does not contain a single component working in isolation but is based on the collaboration of a variety of components which are interacting in well-defined ways. This chapter examines the protocols and standards involved in this topic and investigates to what extent the data required by the EU directive can be stored.

4.1 Requirements of the EU Directive

For Internet e-mail, the EU Data Retention directive 2006/24/EC requires storage of the following information.

Art 5 (1) lit a Z 2 i: "data necessary to trace and identify the source of a communication: ... the user ID(s) allocated"

Art 5 (1) lit a Z 2 ii: "data necessary to trace and identify the source of a communication: ... the user ID and telephone number allocated to any communication entering the public telephone network"

Art 5 (1) lit a Z 2 iii: "data necessary to trace and identify the source of a communication: ... the name and address of the subscriber or registered user to whom an Internet Protocol (IP) address, user ID or telephone number was allocated at the time of the communication"

Art 5 (1) lit b Z 2 ii: "data necessary to identify the destination of a communication: ... the name(s) and address(es) of the subscriber(s) or registered user(s) and user ID of the intended recipient of the communication"

Art 5 (1) lit c Z 2 i: "data necessary to identify the date, time and duration of a communication: ... the date and time of the log-in and log-off of the Internet access service, based on a certain time zone, together with the IP address, whether dynamic or static, allocated by the Internet access service provider to a communication, and the user ID of the subscriber or registered user"

Art 5 (1) lit c Z 2 ii: "data necessary to identify the date, time and duration of a communication: ... the date and time of the log-in and log-off of the Internet e-mail service ... based on a certain time zone"

Art 5 (1) lit d Z 2: "data necessary to identify the type of communication: ...
concerning Internet e-mail ... : the Internet service used"

Art 5 (1) lit e Z 3 i: "data necessary to identify users' communication equipment or what purports to be their equipment: ... the calling telephone number for dial-up access"

Art 5 (1) lit e Z 3 ii: "data necessary to identify users' communication equipment or what purports to be their equipment: ... the digital subscriber line (DSL) or other end point of the originator of the communication"

4.2 Possible Implementation of the EU Directive

The investigation in this thesis focuses on the implications of the EU Data Retention directive including the financial costs and the additional disk space needed. In order to provide an in-depth analysis, a bottom-up approach is chosen: This sections starts by introducing a model mail provider and estimating its network traffic. Furthermore, a data model design is provided and extrapolated to the whole mail provider. Finally, an estimation of the monetary costs caused by the storage and supply of the retained data follows.

4.2.1 Model Mail Provider

In order to quantify the amount of disk space and the costs associated with satisfying the requirements of the EU Data Retention directive for Internet e-mail a fictitious model provider with a certain set of properties is defined.

Data retention period	6 months
Request volume per year	50
Customers	500 000
Received e-mail messages per day and customer	33
Received spam ratio	85%
Sent e-mail messages per day and customer	10

Table 4.1: Model provider figures for Internet e-mail

The model provider introduced here serves 500 000 customers and represents a medium-sized Austrian ISP. By extrapolating the figures in [26], the request volume for Internet e-mail records was assumed to be 50 per year. Following a similar investigation on this topic [26], a ratio of unsolicited bulk and commercial e-mail (UBE and UCE, "spam" – see Section 4.2.4) among all incoming messages was estimated to 85%. A German investigation [27] reported a similar ratio in July 2007.

Concerning estimations for e-mail traffic, the report [26] estimated the average number of incoming / outgoing e-mail messages per day and user

to be 17 / 2 for a French ISP. Another one [58] summarized in Section 1.2 assumed 32 e-mail messages per day and user for a Dutch ISP, without distinguishing between incoming and outgoing messages. Both include spam. Another source [43] investigating corporate environments estimates 99 incoming and 34 outgoing messages per day and user, also including spam. Based on these numbers, an estimation of 33 incoming and 10 outgoing e-mail messages per day and customer was used, including spam, for the model mail provider. Table 4.1 summarizes the figures needed for estimating the disk space.

4.2.2 Storage Requirements

This section is concerned with the costs in terms of needed disk space. Based on the characterization of two different e-mail communication scenarios, the disk space needed for each is evaluated. Further, the probabilities of each scenario to happen are estimated leading to a disk space estimation for the whole web mail provider introduced in Section 3.2.1.

Communication Scenarios

In order to analyze the data processed and therefore retained by the mail provider, a differentiation between two communication setups is necessary.

Scenario 1: Internal Communication (s_1). Scenario one is the simple situation of two customers communicating via the provider's mail server (see Figure 4.1).

Figure 4.1: Scenario 1: Internal communication

The arrows in Figure 4.1 refer to electronic dialogs via POP, IMAP, SMTP, or HTTP (web mail). The same applies to Figure 4.2

Scenario 2 (s_2). Scenario two involves a foreign communication party (see Figure 4.2).
Considering an "outgoing" e-mail message, which refers to a communication traveling away from the mail provider: The customer initiates the whole process by handing a message to its mail server which in turn hands it to its next station, a foreign mail server. At this point, it may either be collected via POP or IMAP or sent along to another mail server via SMTP. The arrows in Figure 4.2 refer to communications in both directions.

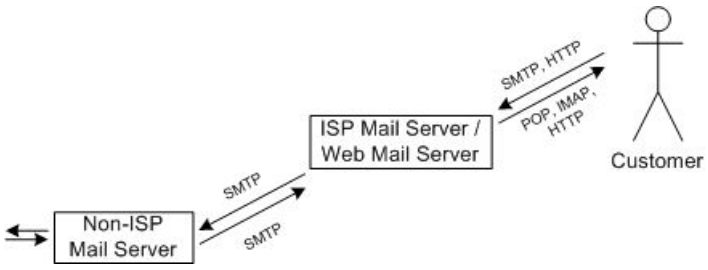

Figure 4.2: Scenario 2: Foreign mail server involved

Scenario two is further divided in two subscenarios depending on the direction of the communication: Incoming ($s_{2/in}$) and outgoing ($s_{2/out}$).

Data Model

For every e-mail processed by the mail provider's mail server, at least one record adhering to the model in Table 4.2 is created. If one message is destined to multiple recipients, one record is written for each recipient.

Field	Type	Bytes needed
ID	Numeric	4
Timestamp	Date	7
Delivery_Date	Date	7
From_E-Mail	Text	1 per character
To_E-Mail	Text	1 per character
From_Customer_Ref	Numeric	4
To_Customer_Ref	Numeric	4

Table 4.2: Storage data model for Internet e-mail

A certain size in bytes is given for each of the fields in Table 4.2. These values are valid for the Oracle database management system[1] (see [37]). Following is a discussion of each of the fields in Table 4.2.

ID This attribute is the so called primary key of the table which uniquely identifies exactly one record. It is not possible for two records to have the same `ID`.

Timestamp This date and time combination field refers to the time when the mail server processed an e-mail for the first time, either by receiving it from a foreign mail server (incoming) or a customer (outgoing).

Delivery_Date This is the time and date when the recipient finally collects an e-mail message from the mail server via the POP or IMAP

[1] http://www.oracle.com/database/index.html

protocols (see Sections 2.4.2 and 2.4.3). Information about the delivery is in general only available for incoming messages. The time and date an outgoing message is collected at a foreign mail server is not known by the sending mail server.

From_E-Mail, To_E-Mail These two fields store the e-mail addresses of the two communication parties. Two electronic sources may be used for collecting this information: The SMTP envelope or the message header.

An e-mail address is composed of a local part and a domain. According to the SMTP specification (RFC2821 [23]), the maximum length is set to 64 characters for the local part and 255 characters for the domain, resulting in a maximum length of 320 characters, including the @-character. For the analysis of the storage requirements, an average length of 50 characters is assumed for these fields.

From_Customer_Ref, To_Customer_Ref These fields are references to the provider's customer database. Table 4.3 shows how this customer database might look like.

ID	**Name**	**Address**	**Telephone number**	···
1	Jane Doe	Address 1	43 1234567	···
2	John Doe	Address 2	43 7654321	···
···	···	···	···	⋱

Table 4.3: Customer database scheme

The identification of the customers by the provider can be done on the basis of the IP address or the mail server authentication. However, except for the special case of both parties being customers of the same ISP, one communication party will always remain unidentified regarding his name, address, etc.

Average Record Size

In order to estimate the disk space requirements for retaining data related to Internet e-mail, an average record size is estimated.

Table 4.4 summarizes to which extent the data required in the EU directive is available in each scenario. The value "x" in a table cell stands for "information available in this scenario" and an empty cell indicates "information not available".

Table 4.4 shows that some of the attributes introduced earlier in this section are not needed in some cases. Therefore, the disk space needed for a record depends on the scenario: 126 bytes for s_1, 122 bytes for $s_{2/in}$, and 115 bytes for $s_{2/out}$.

Attribute \ Scenario	s_1	$s_{2/in}$	$s_{2/out}$
ID	x	x	x
Timestamp	x	x	x
Delivery_Date	x	x	
From_E-Mail	x	x	x
To_E-Mail	x	x	x
From_Customer_Ref	x		x
To_Customer_Ref	x	x	

Table 4.4: Data availability for scenarios

In order to compute an average record size for the Internet access data store, probabilities of occurrence have to be assigned to the Scenarios s_1, $s_{2/in}$, and $s_{2/out}$. Recall that Scenario one refers to an e-mail conversation between two customers of the same mail provider which is probably a relatively rare situation. Therefore, s_1 is assumed to happen with a probability of 20%, the remaining 80% are distributed equally among $s_{2/in}$ and $s_{2/out}$.

These probabilities of occurrence combined with the average record sizes of the scenarios result in an average record size of 120 bytes.

An overall disk space requirement for the entire mail provider is obtained by multiplying the average record size with the e-mail traffic assumptions provided in Section 4.2.1.

Disk Space

The EU Data Retention directive requires the ISPs to store the data for at least six months and for at most two years. Based on the numbers provided for the model mail provider in Section 4.2.1 and this section, it was estimated that an Austrian mail provider serving 500 000 customers with average behavior faces a permanent additional disk space requirement of about 959.8 gigabytes for storing Internet e-mail related data required by the EU directive for a six months period, including a full backup of the stored data.

Consider that with a spam ratio of 85% and the other assumptions made in Section 4.2.1 more than ½ of the required disk space corresponds to useless data.

After having evaluated the needed disk space, the next topic will be the monetary costs caused by an implementation of the EU Data Retention directive.

4.2.3 Costs

In order to provide a standardized view on this topic, the costs described apply to the model mail provider introduced in Section 4.2.1.

The entire task can be subdivided into *data storage* and *data retrieval*. The *data storage* process includes the gathering of traffic and location data, possible transformations applied to the data, and the archiving of the data. *Data retrieval* refers to the information flow from the ISPs to government authorities as a reaction to specific queries.

Whereas the data storage for Internet access related data differs quite much compared to Internet e-mail data, the data retrieval does not. Data retrieval costs are roughly caused by the setup and maintenance of a server and the development of an appropriate user interface which are basically the same for the Internet access and e-mail. Therefore, the data retrieval costs are summarized here and explained in detail in Appendix B.

In terms of personnel costs, the salary charged for one technician was estimated as €120 per hour which corresponds to €19 200 per month per full time equivalent (FTE).

The numbers obtained from [51] and [38] have been rounded up in the last digit.

Data Storage Costs

The storage process includes *gathering*, *processing*, and *archiving* of the required data.

The costs were divided into setup costs incurring once (see Table 4.5) and operational costs incurring every month (see Table 4.6).

Category	Item	Costs [€]	Remarks
HW	storage server	20 090	e.g. Sun StorageTek 5220 (product numbers XTB5220HR10A1-Z, XTB5220HR11A1SB20Z [51])
HW	acquisition server	15 990	e.g. Sun Fire T2000 (product number T20Z108B-16GA2G [51])
DEV	mail server customization	115 200	6 FTEs (2 FTEs for 3 months)
SW	database software	11 160	e.g. Oracle Database Enterprise Edition for one processor [38]
DEV	project setup, software development & deployment	288 000	15 FTEs (2.5 FTEs for 6 months)

Table 4.5: Setup costs for *data storage* of Internet e-mail related data

Category	Item	Costs [€/month]	Remarks
HW	maintenance storage server	110	e.g. Sun StorageTek Gold Support (product number W9D-ST5220-N-24-2G [51])
HW	maintenance acquisition server	160	e.g. Sun Fire Gold Support (product number W9D-T2000-8-24-3G [51])

Table 4.6: Operational costs for *data storage* of Internet e-mail related data

Data Retrieval Costs

It is not enough to store the data to be retained. Provisions need to be made for efficiently accessing the data stored in order to respond to queries. The retrieval process includes *extraction* of the requested data from the data warehouse and *delivering* it to government authorities.

As mentioned before, the data retrieval costs are basically the same for Internet access and Internet e-mail and are therefore summarized in Table 4.7 and explained in detail in Appendix B.

	HW	SW	DEV	GEN	Σ
Setup [€]	15 990	0	115 200	0	131 190
Operation [€/year]	1 920	0	115 200	115 200	232 320

Table 4.7: Data retrieval costs (for more details, see Appendix B)

Total Costs

The data in Table 4.8 shows that the EU Data Retention directive causes overall costs of about €817 190 in the first year, and of about €235 560 in each of the following years for the model mail provider introduced in Section 4.2.1.

	HW	SW	DEV	GEN	Σ
Setup [€]	52 070	11 160	518 400	0	581 630
Operation [€/year]	5 160	0	115 200	115 200	235 560

Table 4.8: Total costs for storage and retrieval for Internet e-mail

4.2.4 Open Issues

The previous sections of this chapter provided suggestions on how the regulations of the EU directive for the area "Internet e-mail" can be im-

plemented. This section provides an overview of the open issues that remain.

Requirements of the EU Directive

Ambiguous formulations and requirements are discussed in the following paragraphs and, additionally, assumptions are made about what could be meant by the authors of the EU directive.

Art 5 (1) lit a Z 2 i-iii: "data necessary to trace and identify the source of a communication: ... the user ID(s) allocated; ... the user ID and telephone number allocated to any communication entering the public telephone network; ... the name and address of the subscriber or registered user to whom an Internet Protocol (IP) address, user ID or telephone number was allocated at the time of the communication"

The formulation of these paragraphs leaves some room for interpretation, at least from a technical perspective. The term "user ID" typically refers to a unique identifier of a certain user. Although this phrase is usually not used in this context and would more apply to Internet telephony, it does make sense to uniquely identify the participants of an e-mail communication. The e-mail address serves as this kind of identifier and is assumed to be the required "user ID" in the EU directive.

Concerning the other terms used in the paragraphs i-iii, it is also unclear whether the required "telephone number allocated to any communication ..." is really meant to apply to Internet e-mail. From a technical perspective, no telephone numbers are allocated to e-mail communications and it is therefore assumed that this requirement applies to Internet telephony (which is not subject of this thesis). Also, the requirement of "the name and address of the subscriber or registered user to whom an Internet Protocol (IP) address, user ID or telephone number was allocated at the time of the communication" is assumed to apply to Internet access and is regarded as not relevant for Internet e-mail (and therefore for the analyzes made in this chapter).

Art 5 (1) lit b Z 2 ii: "data necessary to identify the destination of a communication: ... the name(s) and address(es) of the subscriber(s) or registered user(s) and user ID of the intended recipient of the communication"

The information about the identity of the participants of a communication is not always available to the mail server processing an e-mail message. Consider the case of Scenario 2 introduced in Section 4.2.2, neither does the mail server know who the processed message is handed to by the foreign mail server in case of an outgoing

e-mail, nor does he possess any reliable personal information about the sender in case of an incoming e-mail. The only situation where information about both participants is available is Scenario 1, in which the customers are somehow known to the mail server, for example by an SMTP authorization procedure (see Section 2.4.1) or on the basis of the customer's IP address.

Art 5 (1) lit c Z 2 i: "data necessary to identify the date, time and duration of a communication: ... the date and time of the log-in and log-off of the Internet access service, based on a certain time zone, together with the IP address, whether dynamic or static, allocated by the Internet access service provider to a communication, and the user ID of the subscriber or registered user"

It is not made clear whether this regulation applies to Internet e-mail. The directive literally says "... concerning Internet access, Internet e-mail ... ", but the paragraph below requires certain data concerning "the log-in and log-off of the Internet access service" to be stored. Internet access is technically not connected to Internet e-mail and it is therefore assumed in this thesis that Art 5 (1) lit c Z 2 i does not apply to Internet e-mail.

Art 5 (1) lit c Z 2 ii: "data necessary to identify the date, time and duration of a communication: ... the date and time of the log-in and log-off of the Internet e-mail service ... based on a certain time zone"

The term "Internet e-mail service" used by the recipient of an e-mail is what remains unclear in this paragraph. As the log-in and log-off timestamps are required by the directive it could be referring to the SMTP server used to send or receive an e-mail. The used protocol for this procedure, SMTP, although, is not directly user-guided or involves user-action but is a simple automated transmission of data only initiated by a user and therefore the log-in and log-off time and session duration does not represent any user- or content-dependent information. Therefore the information given is very meaning-limited, see Section 2.4.1 for details on the simple mail transfer protocol (SMTP).

When thinking of web mail providers, there is a log-in procedure in the classical sense, but the required information about the log-in and log-off is available to the mail provider only if the web server is controlled by it. An e-mail originating at a web mail server does not contain any information about when and how long the sender used the web mail interface.

Art 5 (1) lit d Z 2: "data necessary to identify the type of communication: ...
concerning Internet e-mail ... : the Internet service used"

For this requirement, the abstract term "Internet service" is not specified and technically not obvious. It was therefore not accounted for in the implementation proposed in this thesis.

Art 5 (1) lit e Z 3 i-ii: "data necessary to identify users' communication equipment or what purports to be their equipment: ... the calling telephone number for dial-up access; ... the digital subscriber line (DSL) or other end point of the originator of the communication"

The requirements i and ii order the retainment of data concerning the users' end points of communication depending on the equipment used. As it was assumed in this thesis that service providers which exclusively offer Internet e-mail services may also be affected by the EU Data Retention directive, the information required in these paragraphs is not available. Mail providers do not possess any reliable information about their users' communication equipment.

Spam and Botnets

The term "spam", derived from *Spi*ced H*am*, refers to unsolicited bulk and commercial e-mail (UBE and UCE). According to a study from July 2007 more than two out of three e-mail messages transported on the Internet were spam [27]. Vint Cerf, co-developer of the TCP protocol, stated at the World Economic Forum 2007 in Davos that according to his estimations out of the approximately 600 million computers connected to the Internet 150 million of them might be part of botnets [3].

A *botnet* is a group of Internet-connected computers running a special kind of hidden remote-control software without the knowledge of the owners. These botnets are controlled via Internet Relay Chat (IRC) channels by so called botnet herders who can issue commands to thousands of computers instantly and simultaneously. Botnets have become the preferred way for spammers to send out their e-mail messages.

As spam is not excluded in the EU Data Retention directive, this would imply that a big fraction of the incoming messages and associated data retained correspond to useless data. Considering the model mail provider introduced in Section 4.2.1, among all the e-mail messages retained by this provider, a *fraction of 65% corresponds to spam.*

Open Mail Relays

A mail server accepting and transmitting e-mail messages regardless of the source and destination is called an *open mail relay* and his actions are called relaying. The early Internet e-mail system consisted mainly of open mail relays. With the increase of abuse of the Internet e-mail system in the form of spam, mail server operators were forced to take security measures: Nowadays mail servers usually accept e-mail messages either if they are responsible for the source or for the destination. An ISP's

mail server, for example, accepts all e-mail messages being sent from or to customers of this ISP. Nevertheless, there are still open mail relays in the Internet which can be used to transmit messages without the sender's identity being checked by the mail server, thus supporting arbitrary `From` fields in the SMTP envelope and headers, and therefore leading to the storage of false and faked data (see Section 2.4.1 for details on SMTP).

Popular Web Mail Providers and Anonymity

Popular web mail providers include *Yahoo! Mail*[2] which was the most frequently visited website among the US population in march 2007 [18] with $\approx$250 million users, *Microsoft Hotmail*[3] close behind with $\approx$228 million users, and *Google Mail*[4] with $\approx$51 million users [2].

Although registration is required to use each of the services, the identity specified during sign-up is not verified, which is especially interesting in the context of this work. Various data like name, location, gender, and date of birth are asked during the registration procedure for Google Mail, Yahoo! Mail, and GMX[5]. However, setting up a fake account for these freemail providers is a matter of minutes.

[2] http://mail.yahoo.com
[3] http://www.hotmail.com
[4] http://mail.google.com
[5] http://www.gmx.net

4.3 Conclusion

The technical realization of the e-mail system is based on a distributed approach, there is no central authority in control which oversees the global traffic. The EU Data Retention directive requires personal information about sender and recipient to be stored which is not always possible. In addition, some of the formulations in the directive do not make it clear what data exactly has to be stored.

Each e-mail service provider is in control of his own infrastructure and knows his customers, but as soon as data is exchanged with other networks personal information about foreign communication participants is not available. The extent to which the information required by the directive is available depends on the technical circumstances. Generally, the more of the technical infrastructure is controlled by the mail provider, the more reliable data is available. The distributed transmission of e-mail messages as it happens on the Internet, although, leads to the consequence that available data does not equal reliable data, the sender's e-mail address can theoretically be forged in most of the cases for example. In some cases, the required data may not even be available for storage at all.

The overall storage requirement caused by the EU Data Retention directive in the area of Internet e-mail for the model mail provider with 500 000 customers defined in Section 4.2.1 is estimated to be around 959.8 gigabytes (including backup). The monetary costs caused by an implementation are estimated to add up to €817 190 in the first year, and €235 560 in the following years. Additionally, it was estimated that over ½ of the disk space and associated costs is needed for the data retention of unsolicited bulk and commercial e-mail ("spam").

Finally, it has to be pointed out that, despite a full implementation of the regulations in the EU Data Retention directive, technical means are readily available for e-mail communication which remains completely anonymous. This may be done by using non-EU mail providers or fake accounts at popular web mail providers for example.

5 Summary

The Data Retention directive 2006/24/EC of the European Parliament, released on 15.03.2006, requires the operators of publicly accessible electronic communication networks to store and provide traffic and location data generated or processed in their networks to serve the investigation, detection, and prosecution of serious crime. The implementation of this directive is very controversial. In particular, it could lead to discussions on a number of topics, including privacy concerns. This thesis, though, focuses on technical and financial issues related to the implementation of the directive with respect to the guidelines for Internet access and Internet e-mail.

An important question when dealing with the EU Data Retention directive is which data exactly has to be retained by whom. On the one hand, the affected data is not defined in detail, content data is explicitly forbidden to be stored, but unfortunately the border between content and traffic or location data is sometimes blurred in electronic communications. On the other hand, uncertainties are created by general statements that the providers of publicly available services are obliged to retain the data generated or processed by them without relating such requirements to underlying technical aspects. The conclusions made in this thesis are based on the assumption that a single service provider may either implement the EU Data Retention directive guidelines for Internet access or Internet e-mail. An Internet Service Provider (ISP) would therefore be required to only implement the directive with respect to Internet access in case he does not offer any e-mail services. An alternative interpretation would be to assume that Internet access services implicitly enable Internet e-mail.

Concerning *Internet access*, depending on the types of Internet connections provided by an ISP, the data required by the EU Data Retention directive to be stored is to a large extent available. The overall storage requirement caused by the EU directive in the area of Internet access for a model ISP with 500 000 customers is estimated to be around 8.9 gigabytes (including backup). The monetary costs caused by an implementation are estimated to add up to €673 790 in the first year, and €232 320 in the following years.

Concerning *Internet e-mail*, it is important to mention that the e-mail system is based on a distributed approach, there is no central authority in control which oversees the global traffic. Each e-mail service provider is in control of his own infrastructure and knows his customers, but as soon as data is exchanged with other networks, personal information about foreign communication participants is not available. The overall storage require-

ment caused by the EU Data Retention directive in the area of Internet e-mail for a model mail provider with 500 000 customers is estimated to be around 959.8 gigabytes (including backup). The monetary costs caused by an implementation are estimated to add up to €817 190 in the first year, and €235 560 in the following years.
Unsolicited bulk and commercial e-mail (UBE and UCE, "spam") is not explicitly excluded by the EU directive and therefore will use up more than ½ of the overall disk space needed.

Finally, it has to be pointed out that technical means are readily available to remain undetected despite a full implementation of the EU Data Retention directive for Internet access and also for Internet e-mail. For the former, there are easily accessible ways for people with basic technical knowledge to gain Internet access and stay anonymous. For the latter, traffic and location information corresponding to customers of foreign mail service providers which are not subject to the EU directive is not available to the mail providers within the EU.

A Technical Terms

This chapter provides an alphabetically ordered compact description of various technical terms used in this thesis.

ADSL: The **A**synchronous **D**igital **S**ubscriber **L**ine is a technique for digital communication over traditional telephone wires.

Browser: An application used to communicate with web servers, e.g. Mozilla Firefox, Microsoft Internet Explorer, or Opera.

Base64: An algorithm used to encode 8-bit binary data, e.g. executable files, into a character string comprised of only a few different characters.

BCC: Short for **B**lind **C**arbon **C**opy, which is a special type of e-mail delivery. Somebody specified to receive a blind carbon copy of an e-mail will receive a copy of the message, but the other recipients do not notice.

Body: One of the two basic parts of the content of an e-mail, besides the header (see Section 2.4.4).

CC: Short for **C**arbon **C**opy, which is a special type of e-mail delivery. Recipients of carbon copies receive a copy of an e-mail.

Client: A client is a computer demanding services offered by a server.

DSL, xDSL: The **D**igital **S**ubscriber **L**ine (DSL, sometimes xDSL) is a family of techniques for digital communication over traditional telephone wires (see Section 2.2.3).

ESMTP: **E**xtended **SMTP** is an extension mechanism for the Simple Mail Transfer Protocol (SMTP; see Section 2.4.1).

ETSI: The **E**uropean **T**elecommunications **S**tandards **I**nstitute[1] is a European non-profit organization developing telecommunication specifications.

Header: One of the two parts of the content of an e-mail, besides the body (see Section 2.4.4).

Host: Computers on the Internet or networks in general are sometimes referred to as hosts.

[1] http://www.etsi.org

HTML: The **H**ypertext **M**arkup **L**anguage specifies a language used to build websites. Browsers are applications capable of transforming HTML documents into a visual output.

HTTP: The **H**yper**t**ext **T**ransfer **P**rotocol is a protocol used by web browsers and web servers for communication with each other.

IEEE: **I**nstitute of **E**lectrical and **E**lectronics **E**ngineers[2], an international non-profit organization, which, for example, develops specifications for communication on the Internet.

IETF: **I**nternet **E**ngineering **T**ask **F**orce[3], a non-profit organization developing improvements for the Internet architecture.

IMAP: **I**nternet **M**essage **A**ccess **P**rotocol, an advanced protocol to access and manipulate electronic mailboxes (see Section 2.4.3).

IMF: The content of an e-mail is built as specified by the **I**nternet **M**essage **F**ormat.

IP: The **I**nternet **P**rotocol is the most used network-layer protocol (see Section 2.3.1).

IP Address: An address used to deliver packets to a particular computer in an IP network (see Section 2.3.1).

IRC: **I**nternet **R**elay **C**hat, a popular chat protocol which integrates multiple servers into a large communication network.

ISO: **I**nternational **O**rganization for **S**tandardization, an international non-profit organization developing specifications for a broad range of topics.

ISP: **I**nternet **S**ervice **P**rovider, a company offering Internet connections and various other electronic services like e-mail access.

LAN: A **L**ocal **A**rea **N**etwork is a network of computers covering a small geographic area.

Login, Logout: The process of starting or ending to use a certain electronic service. The login most often is done together with a form of authentication, for example, by providing a username / password pair.

MAC Address: Each piece of networking hardware cen be uniquely identified by its **M**edia **A**ccess **C**ontrol address.

[2] http://www.ieee.org
[3] http://www.ietf.org

MD5: **M**essage-**D**igest algorithm **5** is a popular hash-function. Hash-functions are mathematical algorithms transforming a certain input into an output in such a way that makes it very hard to compute the input based on the output. MD5 transforms every input, regardless of the size, into a 128-bit output.

Modem: **Mo**dulator **Dem**odulator, a networking device which is able to convert analog signals to digital ones and vice-versa.

MIME: **M**ultipurpose **I**nternet **M**ail **E**xtensions, an e-mail extension specifying ways of embedding non-text data into e-mail messages.

MTA: The **M**ail **T**ransfer **A**gent is a software application responsible for accepting, transferring, and organizing e-mail messages. Computers operating this software are referred to as mail servers.

MUA: **M**ail **U**ser **A**gent, an application used to hand outgoing mails to a mail server and fetch incoming mails from the same. Microsoft Outlook and Mozilla Thunderbird are popular examples of mail user agents.

OSI / ISO: The **O**pen **S**ystems **I**nterconnection is an effort to standardize networking, initiated by the **I**nternational **O**rganization for **S**tandardization (ISO).

POP: **P**ost **O**ffice **P**rotocol, a protocol to access and manipulate electronic mailboxes (see Section 2.4.2).

Protocol: A protocol in the context of electronic communication is a well-defined procedure for the exchange of information between computers. Protocols are formalized in so called standards or specifications.

RFC: **R**equests **F**or **C**omments are documents describing new protocols or methods for communication on the Internet which may be adopted as official standards by the Internet Engineering Task Force (IETF)[4].

Router: A piece of networking hardware placed at the interface between two networks which is responsible for forwarding packets from one to the other.

SASL: The **S**imple **A**uthentication and **S**ecurity **L**ayer is a protocol-independent framework for authentication and data security used by many application protocols on the Internet.

Server: A server is a computer offering services to a client.

Session: Two-way protocol dialogs are often referred to as sessions. E-mail messages are exchanged in SMTP sessions.

[4] http://www.ietf.org

Spam: The term spam refers to unsolicited bulk and commercial e-mail (UBE and UCE).

SMTP: The **S**imple **M**ail **T**ransfer **P**rotocol is used to transfer mails from MUA to MTA and from MTA to MTA (see Section 2.4.1).

SMTP AUTH: An extension to the SMTP protocol which provides authentication of the SMTP client (see Section 2.4.1).

SSL: **S**ecure **S**ockets **L**ayer, a cryptographic protocol used to secure various communication services such as e-mail or web browsing.

TLS: **T**ransport **L**ayer **S**ecurity, the successor of SSL.

TCP/IP: The **T**ransmission **C**ontrol **P**rotocol is a connection-oriented protocol on top of IP. It is the standard protocol used on the Internet, therefore often referred to as TCP/IP (see Section 2.3.1).

UDP: **U**ser **D**atagram **P**rotocol, a connection-less protocol on top of IP.

User(-name, -ID): Unique identifier for users of a certain system, a voice-over-IP (VoIP) application for example.

Web mail: A website providing access to a user's mailbox is called "web mail interface" (see Section 2.4.6).

Web server: A computer serving HTML pages to browsers via HTTP.

WEP: **W**ireless **E**quivalent **P**rivacy, an encryption method for W-LAN communication.

W-LAN: **W**ireless **L**ocal **A**rea **N**etwork, a wireless communication network.

B Data Retrieval Costs

As mentioned in Section 3.2.1, 200 requests by government authorities per year were estimated in [26] for Internet access. In the same manner, as mentioned in Section 4.2.1, 50 requests by government authorities per year were estimated in [26] for Internet e-mail. Concerning the data retrieval costs, ½ FTE was considered enough to handle this request load.

The numbers obtained from [51] and [38] have been rounded up in the last digit.

Category	**Item**	**Costs [€]**	**Remarks**
HW	data access server	16 000	e.g. Sun Fire T2000 Server (Product Number T20Z108B-16GA2G [51])
DEV	development internal interface	115 200	6 FTEs (1 FTE for 6 months)

Table B.1: Setup costs for *data retrieval* of Internet access and e-mail

Category	**Item**	**Costs [€/month]**	**Remarks**
HW	maintenance access server	200	e.g. Sun Gold Support (product number W9D-T2000-8-24-3G [51])
DEV	interface maintenance	9 600	0.5 FTEs
GEN	request handling	9 600	0.5 FTEs

Table B.2: Operational costs for *data retrieval* of Internet access and e-mail

List of Figures

List of Tables

Bibliography

[1] A1. A1 Hotspots. http://www.a1.net/hotspots.

[2] M. Arrington. A Comparison of Live Hotmail, Gmail and Yahoo Mail, Feb. 2007. http://www.techcrunch.com/2007/02/08/a-comparison-of-live-hotmail-gmail-and-yahoo-mail.

[3] BBC. Criminals 'may overwhelm the web', Jan. 2007. http://news.bbc.co.uk/1/hi/business/6298641.stm.

[4] I. A. Board and L. Chapin. Applicability Statement for OSPF. RFC 1370 (Historic), Oct. 1992.

[5] D. Boneh. Twenty Years of Attacks on the RSA Cryptosystem, Feb. 1999.

[6] D. Crocker. STANDARD FOR THE FORMAT OF ARPA INTERNET TEXT MESSAGES. RFC 822 (Standard), Aug. 1982. Obsoleted by RFC 2822, updated by RFCs 1123, 2156, 1327, 1138, 1148.

[7] S. Deering and R. Hinden. Internet Protocol, Version 6 (IPv6) Specification. RFC 2460 (Draft Standard), Dec. 1998.

[8] W. Diffie and M. Hellman. Special Feature Exhaustive Cryptanalysis of the NBS Data Encryption Standard, June 1977.

[9] DSL Forum. DSL Forum. http://www.dslforum.org/learndsl/aboutadsl.shtml.

[10] Electronic Frontier Foundation. Cracking DES - Secrets of Encryption Research, Wiretap Politics & Chip Design, July 1998.

[11] European Parliament and the Council of the European Union. Directive 2006/24/EC of the European Parliament and of the Council, Mar. 2006. http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/l_105/l_10520060413en00540063.pdf.

[12] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. RFC 2616 (Draft Standard), June 1999. Updated by RFC 2817.

[13] S. Fluhrer, I. Mantin, and A. Shamir. Attacks on RC4 and WEP, 2002. http://www.wisdom.weizmann.ac.il/~itsik/RC4/Papers/rc4_wep.ps.

[14] N. Freed and N. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples. RFC 2049 (Draft Standard), Nov. 1996.

[15] N. Freed and N. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045 (Draft Standard), Nov. 1996. Updated by RFCs 2184, 2231.

[16] N. Freed and N. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. RFC 2046 (Draft Standard), Nov. 1996. Updated by RFCs 2646, 3798.

[17] N. Freed, J. Klensin, and J. Postel. Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures. RFC 2048 (Best Current Practice), Nov. 1996. Obsoleted by RFCs 4288, 4289, updated by RFC 3023.

[18] Hitwise. US - Top 20 Websites March 2007. `http://www.hitwise.com/datacenter/rankings.php`.

[19] P. Hoffman. SMTP Service Extension for Secure SMTP over TLS. RFC 2487 (Proposed Standard), Jan. 1999. Obsoleted by RFC 3207.

[20] JiWire. WiFi HotStats as of March 2007. `http://www.jiwire.com/search-hotspot-locations.htm`.

[21] JiWire. Worldwide Wi-Fi Hotspots Hits the 100,000 Mark, Jan. 2006. `http://www.jiwire.com/press-100k-hotspots.htm`.

[22] S. Josefsson. The Base16, Base32, and Base64 Data Encodings. RFC 4648 (Proposed Standard), Oct. 2006.

[23] J. Klensin. Simple Mail Transfer Protocol. RFC 2821 (Proposed Standard), Apr. 2001.

[24] J. Klensin, N. Freed, M. Rose, E. Stefferud, and D. Crocker. SMTP Service Extensions. RFC 1869 (Standard), Nov. 1995. Obsoleted by RFC 2821.

[25] G. Klyne, M. Nottingham, and J. Mogul. Registration Procedures for Message Header Fields. RFC 3864 (Best Current Practice), Sept. 2004.

[26] Marpij, Insight, Boivin, and Associés. Evaluation of the economic impacts of the data retention obligations relating to electronic communications, Sept. 2006.

[27] MessageLabs. MessageLabs Intelligence: Juli 2007. `http://de.messagelabs.com/mlireport/MLI_July2007_DE.pdf`.

[28] P. Mockapetris. Domain names - concepts and facilities. RFC 1034 (Standard), Nov. 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592.

[29] P. Mockapetris. Domain names - implementation and specification. RFC 1035 (Standard), Nov. 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 1995, 1996, 2065, 2136, 2181, 2137, 2308, 2535, 2845, 3425, 3658, 4033, 4034, 4035, 4343.

[30] K. Moore. MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text. RFC 2047 (Draft Standard), Nov. 1996. Updated by RFCs 2184, 2231.

[31] J. Myers. Simple Authentication and Security Layer (SASL). RFC 2222 (Proposed Standard), Oct. 1997. Obsoleted by RFCs 4422, 4752, updated by RFC 2444.

[32] J. Myers. SMTP Service Extension for Authentication. RFC 2554 (Proposed Standard), Mar. 1999. Obsoleted by RFC 4954.

[33] J. Myers and M. Rose. Post Office Protocol - Version 3. RFC 1939 (Standard), May 1996. Updated by RFCs 1957, 2449.

[34] Network Working Group. RFC.net repository of RFC documents. `http://rfc.net`.

[35] C. Newman. Using TLS with IMAP, POP3 and ACAP. RFC 2595 (Proposed Standard), June 1999. Updated by RFC 4616.

[36] NRC Handelsblad. Judicial authorities want to know a lot about calling behavior, Sept. 2005. `http://www.xs4all.nl/uk/overxs4all/privacy/privacy_jaarverslag2005.php#id15`.

[37] Oracle Corporation. Oracle Database SQL Reference 10g Release 2 (10.2) Part Number B14200-02. `http://download.oracle.com/docs/cd/B19306_01/server.102/b14200/toc.htm`.

[38] Oracle Corporation. Oracle Technology Global Price List, Aug. 2007. `http://www.oracle.com/corporate/pricing/technology-price-list.pdf`.

[39] J. Palme. Common Internet E-Mail Headers. `http://people.dsv.su.se/~jpalme/ietf/mail-headers/`.

[40] J. Postel. Internet Protocol. RFC 791 (Standard), Sept. 1981. Updated by RFC 1349.

[41] J. Postel. Transmission Control Protocol. RFC 793 (Standard), Sept. 1981. Updated by RFC 3168.

[42] J. Postel. Simple Mail Transfer Protocol. RFC 821 (Standard), Aug. 1982. Obsoleted by RFC 2821.

[43] Radicati Group. Email archiving corporate survey, 2004-2005. http://www.radicati.com/uploaded_files/news/EA_SurveyPR.pdf.

[44] D. Raggett, A. L. Hors, and I. Jacobs. Hypertext Markup Language (HTML) 4.01 specification, Dec. 1999. http://www.w3.org/TR/html4/.

[45] P. Resnick. Internet Message Format. RFC 2822 (Proposed Standard), Apr. 2001.

[46] R. Rivest. The RC4 Encryption Algorithm, Mar. 1992.

[47] R. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems, 1978.

[48] Rundfunk und Telekom Regulierungs-GmbH (RTR). Austrian Numbering Plan, May 2004. http://www.rtr.at/web.nsf/englisch/Telekommunikation_Nummerierung_Nationale+Nummern_nationaleRufnummern_E129.

[49] SelfNet. ISP SelfNet. http://www.selfnet.at.

[50] Statistik Austria. IKT-Einsatz 2006. http://www.statistik.at/dynamic/wcmsprod/idcplg?IdcService=GET_NATIVE_FILE&dID=48168&dDocName=019136.

[51] Sun Microsystems, Inc. Sun Enduser Price List U.S., Aug. 2007. http://blogs.sun.com/marler/resource/price_lists/USD_MASTER-pricelist_US.pdf.

[52] T-Mobile. T-Mobile Hotspots. http://www.t-mobile.at/business/mobiles_arbeiten/mobiles_internet/hotspot/index.html.

[53] U.S. Department of Commerce/National Institute of Standards and Technology. Digital Signature Standard (DSS), May 1994. http://www.itl.nist.gov/fipspubs/fip186.htm.

[54] U.S. Department of Commerce/National Institute of Standards and Technology. Data Encryption Standard (DES), Oct. 1999. http://csrc.nist.gov/publications/fips/fips46-3/fips46-3.pdf.

[55] U.S. Department of Commerce/National Institute of Standards and Technology. Advanced Encryption Standard (AES), Nov. 2001. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.

[56] U.S. Department of Commerce/National Institute of Standards and Technology. Recommendation for the Triple Data Encryption Algorithm (TDEA) Block Cipher, May 2004. `http://csrc.nist.gov/publications/nistpubs/800-67/SP800-67.pdf`.

[57] US National Bureau of Standards. Data Encryption Standard (DES), Jan. 1977.

[58] Verdonck, Klooster, and Associates. Study into the national implementation of the European data retention directive, Oct. 2006.

[59] H. Wu. The Misuse of RC4 in Microsoft Word and Excel, 2005.