Implications of the EU Data Retention Directive 2006/24/EC

G. Stampfel, W. N. Gansterer, M. Ilger

Abstract: The strongly discussed EU Data Retention directive 2006/24/EC released in 2006 requires the operators of publicly accessible electronic communication networks to retain traffic and location data for various services provided in order to serve the investigation, detection, and prosecution of serious crime. In this paper, we focus on the regulations of the directive in the areas of Internet access, Internet e-mail, and Internet telephony. We analyze how the requirements of the directive can be satisfied and to what extent it is possible to achieve its objectives with the technical means currently available. Moreover, we estimate the resulting costs.

1 Introduction

The *Data Retention directive* 2006/24/EC of the European Parliament [Eur06], published in 2006, requires the operators of publicly accessible electronic communication networks to store ("retain") certain data which is generated or processed in their networks to serve the investigation, detection, and prosecution of serious crime. Affected are traffic and location data (but not the contents of the communication) in various areas: fixed network telephony, mobile telephony, as well as Internet access, Internet e-mail, and Internet telephony for a time period between six months and two years.

In this paper, the focus is on technological aspects of the latter three areas, namely, Internet access, Internet e-mail, and Internet telephony. The objective is to analyze *what* can be done, with the technical means currently available, to satisfy the requirements of the EU directive and *how* this can be achieved. Various statements have been made about the resulting costs. However, many of them are somewhat elusive, because the underlying assumptions and methods used for deducing them are not rigorously documented. We make an attempt to develop a fully transparent and parametrizable model for implementing the Data Retention directive in the areas Internet access, Internet e-mail, and Internet telephony and for estimating the costs arising in these three areas as accurately as possible. Various other very important topics in the context of EU Data Retention directive, such as privacy, integrity of the data retained, confidentiality of access to data retained, or availability of the data retention system require extensive discussions on their own and thus go beyond the scope of this paper.

After a brief summary of related investigations, we review the original text of the EU directive. For the purpose of the cost analysis, a fictitious model provider is introduced. A detailed discussion of how the directive can be implemented and the open issues that remain for each of the three areas of interest follows. Discussions of the implementation

include a review of the technical background and a detailed proposal of a data model for the data to be stored. Finally, a cost estimation covering storage as well as monetary costs associated with the implementation is provided.

Related Reports. An investigation conducted in the Netherlands [Ver06] estimated the costs to vary between €133 million and €157 million and 365 terabytes of disk space for a retention period of one year for the entire Dutch market with about 5 million Internet connected households in 2005. These figures are based on the extrapolation of a model provider with costs of roughly €1.03 per subscriber per month. Various implementation options are assessed in [Ver06]. *Centralized* storage with *direct* and *automated* access for the government authorities to the data store was identified as the best implementation for the Dutch telecommunications market. According to [Ver06], this solution is expected to generate expenses of about €133 million for the entire Dutch market for a period of five years. In contrast to the approach pursued in [Ver06], the work summarized here focusses on *decentralized* storage of the information to be retained.

For a provider with one million customers, a French report [Mar06] estimates up to 12 300 terabytes of data and total annual costs of over €132 million for a data retention period of one year. This corresponds to costs of roughly €11 per subscriber per month.

Analogously to the approach pursued in this paper, the estimates provided in [Mar06] assume a decentralized solution without direct access by the government authorities. However, in contrast to this paper, one important focus of [Mar06] is on legal aspects: Based on the current legal situation in France, which partly contradicts the regulations of the EU directive, it addresses aspects related to the anonymization of data as well as related to *who* has to collect the data to be retained. Moreover, the estimates of costs and storage requirements in [Mar06] are very high and the assumptions on traffic volume are not necessarily representative for the Austrian situation. As a consequence, the central objective of the effort summarized here was to put more emphasis on the technical aspects and to develop a (potentially more accurate) parameterized model for implementing the regulations of the EU directive, which maps those regulations as directly as possible on the technical infrastructure. Based on this mapping, the minimal resource requirements in terms of storage and costs can be identified using regionally accurate traffic volume estimations.

1.1 Requirements of the EU Directive

The basic requirements of the original text of the EU Data Retention directive [Eur06] in the areas Internet access, Internet e-mail, and Internet telephony are as follows.

Art 5 (1) lit a Z 2: "... data necessary to trace and identify the source of a communication: ... concerning Internet access, Internet e-mail and Internet telephony: ... 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"

Art 5 (1) lit b Z 2: "... data necessary to identify the destination of a communication: ... concerning Internet e-mail and Internet telephony: ... the user ID or telephone number of the intended recipient(s) of an Internet telephony call, ... 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: "... data necessary to identify the date, time and duration of a communication: ... concerning Internet access, Internet e-mail and Internet telephony: ... 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, ... 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"

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

Art 5 (1) lit e Z 3: "... data necessary to identify users' communication equipment ... concerning Internet access, Internet e-mail and Internet telephony: ... the calling telephone number for dial-up access, ... the digital subscriber line (DSL) or other end point of the originator of the communication"

1.2 A Model Provider

In order to quantify the amount of disk space and the costs associated with an implementation of the EU directive a fictitious model provider is defined and referred to throughout this paper. We assume that the model provider serves 500 000 customers and has to respond to 300 data requests of government authorities per month.

Assumptions for Internet Access. According to [Sta06], 63.2% of the Austrian households are connected to the Internet via broadband and the rest via dial-up. For the model provider this results in 316 000 broadband accounts. We assumed two dial-up logins / one broadband login per day and customer.

Assumptions for Internet E-Mail. Many estimates of the fraction of unsolicited bulk and commercial e-mail (UBE and UCE,"spam") among all incoming messages are available (see, for example, [Mar06, Mes07]), and they often differ significantly. Nevertheless, there is common agreement that this fraction continues to be at a very high level. For the purpose of this paper, we conservatively estimated the fraction of spam in incoming e-mail 85%.

[Mar06] estimates the average number of incoming / outgoing e-mail messages per day and user to be 17 / 2 for a French ISP. [Ver06] assumes 32 e-mail messages per day and user for a Dutch ISP, without distinguishing between incoming and outgoing messages. Both numbers include spam. [Rad04] investigates corporate environments and estimates 99 incoming and 34 outgoing messages per day and user, also including spam. However, these numbers are not necessarily representative for the Austrian situation. Based on averaging the information provided to us by representative Austrian e-mail providers, we estimate

on average 33 incoming and 10 outgoing e-mail messages per day and mailbox, which includes a spam percentage in the incoming e-mail of 85%.

Assumptions for Internet Telephony. Based on [Mar06], an average customer is estimated to initiate four and receive three Internet telephony calls each day.

2 Internet Access

Various types of Internet connections are available. They slightly vary in their technical details and consequently in the aspects relevant from the point of view of the EU directive.

2.1 Implementation of Data Retention

For every Internet access, one record comprising the attributes *ID*, *Customer_Ref*, *Connection*, *Log-in*, *Log-off*, *IP*, and *Source_Telephone_Number* is stored. The attribute *ID* (numeric, 4 bytes[1]) is the primary key of the table which uniquely identifies exactly one record. Concerning *Customer_Ref* (numeric, 4 bytes) we assume the existence of an ISP-internal customer database. It is not necessary to store name and address twice, therefore, only a reference is used. *Connection* (numeric, 4 bytes) tells whether the connection was established through dial-up, DSL, cable modem, etc.. Time stamps for the date, time, and time zone of the moment when the user established and released his connection to the Internet are contained in the attributes *Log-in* and *Log-off* (timestamp, 7 bytes each). *IP* (numeric, 4 bytes) holds the public IP address assigned to the user's hardware. The directive's requirement to store the originating endpoint is met by the attribute *Source_Telephone_Number* (15 bytes are assumed to be sufficient based on [Run04]).

For currently popular methods for Internet access such as narrowband dial-up, Integrated Services Digital Network (ISDN), Digital Subscriber Line (DSL), Cable, and Wireless LAN (W-LAN), etc., the data required by the EU directive is to a large extent available. Usually, customers also have to register with their names and addresses. The provider is therefore able to assign a customer reference (attribute *Customer_Ref*) to each Internet access session. The same is true for the IP address assigned to the customer's networking hardware (attribute *IP*), as well as the originating telephone number in case of dial-up access and ISDN (attribute *Source_Telephone_Number*). For DSL and other locally tied broadband connections, the originating endpoint should be the customer's home address.

[1] The byte values used in this paper are consistent with the Oracle database management system (see [Ora]).

2.2 Open Issues

Some open issues remain in the area of Internet access which leaves space for circumventing the purpose of data retention. In particular, some formulations and requirements in the text of the EU directive are unclear or ambiguous from a technical point of view. For example, in Art 5 (1) lit a Z 2 the term "user ID" is used. It normally refers to a unique identifier of a certain user, a concept which is typically not needed for Internet access. Moreover, only in the case of dial-up access a telephone number can be assigned to Internet access. These formulations are most likely intended to apply to Internet telephony.

Even with data retention strategies in place, not all scenarios of Internet access can be tracked. There are several situations in which an ISP does not possess any personal data about the customer he is serving, such as free wireless LAN hotspots (for example, at a café) or in the case of dial-up providers without authentication (for example, the Austrian ISP SelfNet (`http://www.selfnet.at`). In these cases, only fairly limited information about the equipment used to connect to the provider or the telephone number can be retained, but no identification of a person.

3 Internet E-Mail

The implementation of the global e-mail system is based on a stack of different protocols, and several different technical setups have to be considered. In our proposal for an implementation of the EU Data Retention directive for this area we distinguish four specific scenarios and their implications for data retention.

3.1 Implementation of Data Retention

An Internet service provider has basically two options for monitoring the e-mail traffic in his network and the e-mail exchange with other networks: Either the packets are processed by an ISP-controlled mail server or they are observed at the network boundary.

The most popular protocols used for transferring e-mail messages are the Simple Mail Transfer Protocol (SMTP, as defined in [Kle01]), the Post Office Protocol (POP, as defined in [MR96]), and the Internet Message Access Protocol (IMAP, as defined in [Cri03]). SMTP specifies how a message is transmitted to its destination mail server, whereas POP and IMAP specify how it is collected from there. The format of an e-mail message is specified by the Internet Message Format (IMF) in [Res01]: Each e-mail consists of an envelope and a content section. The envelope is used to correctly deliver the content, which in turn is subdivided into the header and the body. Especially interesting in the context of the EU directive are the header fields *From* and *To*, because they contain information about sender and recipient of an e-mail. However, messages may contain more or less arbitrary header information and still may reach their destination, as this information can be faked and the transfer process is based on the information exchanged in the SMTP dialogue,

which is not necessarily cross-checked with the header information. Mail servers are not even allowed to reject an e-mail based on header information [Kle01].

Another important way of accessing e-mail functionality is via *web mail*. In this context, a relevant protocol is the Hypertext Transfer Protocol (HTTP) protocol (as defined in [FGM+99]). All the interaction with a mailbox happens through a web browser using the HTTP protocol, and therefore the information to be retained according to the EU Data Retention directive would have to be extracted from HTTP traffic, which is extremely costly. Moreover, another question is raised. Using web mail makes it possible to utilize e-mail services via an ISP who only provides Internet access. Based on the original text of the EU directive it is unclear whether such an ISP should be required to retain data related to Internet e-mail.

Data Model. In order to implement the EU directive, for every e-mail observed by the ISP, one record comprising the attributes *ID*, *Timestamp*, *Delivery_Date*, *From_E-Mail*, *To_E-Mail*, *From_Customer_Ref*, *To_Customer_Ref*, *From_IAccess_Ref*, and *To_IAccess_Ref* could be stored: The attribute *ID* (numeric, 4 bytes) is the primary key of the table which uniquely identifies exactly one record. *Timestamp* (time stamp, 7 bytes) refers to the time when the ISP observes an e-mail message, either when an ISP's mail server handles an e-mail or when a network analyzer observes the SMTP packets. *Delivery_Date* (time stamp, 7 bytes) is the time and date when the recipient finally collects an e-mail message from the mail server. The fields *From_E-Mail* and *To_E-Mail* (text, average length of 50 bytes each) store the e-mail addresses of the two communicating parties. *From_Customer_Ref* and *To_Customer_Ref* (numeric, 4 bytes each) are references to the provider's customer database. The attributes *From_IAccess_Ref* and *To_IAccess_Ref* (numeric, 4 bytes each) are references to entries in the Internet access data model (see Section 2.1).

Data identifying an e-mail communication party with respect to his name, address, etc., is in general only available to an ISP if the user is one of its customers. He is then able to assign a customer record (attribute *Customer_Ref*) to this party, for example, based on the assigned IP address or the mail server authentication (as defined in [SM07]). Otherwise, the only information available about a participant in e-mail communication is the e-mail address from the message header which may be forged as mentioned before. Also, data about the user's Internet access (attributes *From_IAccess_Ref* and *To_IAccess_Ref*) is only available if he is a customer of the ISP concerned. The date and time of the log-in and log-off from the Internet e-mail service, which can basically be interpreted a the time of message transmission and message reception (attribute *Delivery_Date*), is generally available only if the corresponding POP or IMAP packets are somehow observable by the ISP. The SMTP envelope or the header may be taken as sources for the addresses of sender and recipient (attributes *From_E-Mail* and *To_E-Mail*) with the restrictions mentioned above.

Communication Scenarios. In order to cover the most important technical setups, we distinguish four different scenarios.

In *Scenario one* ("ISP Mail Server Only", s_1) two customers exchange e-mail messages using the ISP's mail server. The customers' Internet access is left unspecified, the mail server may also be used from outside the ISP's network, schematically: Customer 1 $\leftrightarrow$ ISP's mail server $\leftrightarrow$ Customer 2. The communication between the mail server and the

customers may be based on SMTP, POP, IMAP, or HTTP (web mail).

In *Scenario two* ("ISP/Non-ISP Mail Server Mix", s_2), the ISP's own mail server is used again. This time, however, the communication is not between two customers of this ISP but between a customer and a foreign party, schematically: Foreign party $\leftrightarrow$ Non-ISP mail server $\leftrightarrow$ ISP's mail server $\leftrightarrow$ Customer. The communication between the mail server and the customer may be based on SMTP, POP, IMAP, or HTTP (web mail).

In *Scenario three* ("Non-ISP Mail Server", s_3) a foreign mail server, outside the ISP's network, is used for e-mail transmission and reception, schematically: Foreign party $\leftrightarrow$ Non-ISP mail server $\leftrightarrow$ Customer. The communication between the mail server and the customer may be based on SMTP, POP, or IMAP (HTTP is considered separately in Scenario four). The customer is inside the ISP's network but uses an external mail server. Therefore, the customer's e-mail traffic is visible to the ISP only on the network boundary. Scenario three is of particular interest because it represents the only scenario where encryption, which is possibly applied to the communication, does make a difference. "Encryption" in this case refers to an encrypted dialog of the concerned protocol as it is specified in RFC3207 [Hof02] for SMTP and in RFC2595 [New99] for POP and IMAP. An ISP cannot be expected to be able break such an encryption for obvious legal and technical reasons.

In *Scenario four* ("Non-ISP Web Mail", s_4) a web mail service which is not controlled by the ISP is used by the customer to send and receive e-mail messages, schematically: Foreign party $\leftrightarrow$ Non-ISP Web mail server $\leftrightarrow$ Customer. The communication between the mail server and the customer is based only on HTTP (web mail). According to the EU Data Retention directive only traffic and location data is to be retained [Eur06]. In this scenario, though, the e-mail-related data is embedded in web traffic (HTTP packets) and can be considered *content* of the communication, thus not to be retained.

Subscenarios. In order to further analyze the different situations of e-mail communication and their implications for data retention, subscenarios have to be distinguished.

Scenario one can be subdivided into four subscenarios, depending on how many customers are inside the provider's network: Two customers inside (Subscenario 1 of Scenario 1, $s_{1/1}$), one customer as receiving party inside ($s_{1/2}$), one customer as sending party inside ($s_{1/3}$), and no customer inside ($s_{1/4}$).

Scenario two also contains four cases, depending on the location of the customer and the direction of the communication: Customer inside and direction incoming (Subscenario 1 of Scenario 2, $s_{2/1}$), customer inside and direction outgoing ($s_{2/2}$), customer outside and direction incoming ($s_{2/3}$), and customer outside and direction outgoing ($s_{2/4}$).

In Scenario three, the direction of the e-mail and the encryption are relevant: Incoming e-mail (Subscenario 1 of Scenario 3, $s_{3/1}$) and outgoing e-mail ($s_{3/2}$). In the case of encrypted communication ($s_{3/3}$), the ISP is unable to identify the packets as SMTP traffic.

Scenario four is not visible at all to the ISP and is therefore not further dividable.

Average Record Size. With the definition of the communication scenarios for Internet e-mail and the data model introduced in Section 3.1, we are able to compute the corresponding record sizes (see Table 1).

$\bar{b}_{1/1}$	$\bar{b}_{1/2}$	$\bar{b}_{1/3}$	$\bar{b}_{1/4}$	$\bar{b}_{2/1}$	$\bar{b}_{2/2}$	$\bar{b}_{2/3}$	$\bar{b}_{2/4}$	$\bar{b}_{3/1}$	$\bar{b}_{3/2}$	$\bar{b}_{3/3}$	$\bar{b}_4$
134	130	130	126	126	119	122	115	126	119	0	0

Table 1: Average record size for scenarios in bytes, $\bar{b}_{x/y}$ denotes the subscenario y of scenario x

We assume that the four scenarios are all equally likely. The subscenarios inside a scenario are also assumed to be equally distributed (see Table 2, recall that Scenario three is roughly divided into an unencrypted ($p_{3/1} + p_{3/2}$) and an encrypted variant ($p_{3/3}$) with each option being equally likely). The probabilities shown in Table 2 combined with the record sizes

$p_{1/1}$	$p_{1/2}$	$p_{1/3}$	$p_{1/4}$	$p_{2/1}$	$p_{2/2}$	$p_{2/3}$	$p_{2/4}$	$p_{3/1}$	$p_{3/2}$	$p_{3/3}$	p_4
6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	12.5	25

Table 2: Probability of occurrence for scenarios in percent, $p_{x/y}$ denotes the subscenario y of scenario x

of each scenario shown in Table 1 result in an overall average record size of 78 bytes.

3.2 Open Issues

Requirements of the EU Directive. Some formulations and requirements in the text of the EU directive are unclear or ambiguous from a technical point of view. Relating to **Art 5 (1) lit a Z 2 iii, Art 5 (1) lit b Z 2 ii**, data identifying the source or destination of an e-mail communication with respect to name, address, etc., is in general only available to an ISP if he is providing Internet access to the respective party. Relating to **Art 5 (1) lit c Z 2 ii**, the required time and date at which a message is collected or handed in by a user is not always known. For example, the date at which an outgoing message is collected at a foreign mail server is not available to an ISP.

Spam. According to [Mes07], in July 2007 more than two out of three e-mail messages transported on the Internet were spam. Thus, a big portion of the storage used for data retention (based on our assumptions, almost 60% !) would be occupied by useless data.

Bypassing Data Retention. There are at least two important technological aspects which allow for circumventing the purpose of the Data Retention directive. With *web mail*, traffic and location data are embedded into the bodies of HTTP requests and responses. Consequently, this information becomes *content* of an electronic communication and, according to the directive, must not be stored. Another issue related to web mail is that various companies (for example, Yahoo or GMX) offer free accounts which do not require any form of identification for registration. It therefore becomes very easy to communicate via e-mail without being identified [Sta07].

The other aspect relates to mail servers which are *external* to the scope of an ISP or of the EU Data Retention directive. The mail server in Scenario three above is not controlled

by the ISP, and therefore from the point of view of this ISP the only source for e-mail data is the traffic exchanged at the network boundary. If these packets are encrypted, for example, using the concepts explained in [Hof02] and [New99], the ISP is left without any data about this communication as he can in general not be expected to break the encryption.

4 Internet Telephony

Internet telephony, also called Voice-over-IP (VoIP), is a relatively recent development compated to Public Switched Telephone Network (PSTN), the traditional landline network, and Public Land Mobile Network (PLMN), and the wireless telephone networks.

4.1 Implementation of Data Retention

For every VoIP conversation observed by the ISP, one record comprised of the attributes *ID*, *From_IAccess_Ref*, *To_IAccess_Ref*, *From_ID_Number*, *To_ID_Number*, *Start_Time*, and *End_Time* is stored: The attribute *ID* (numeric, 4 bytes) is the so called primary key of the table which uniquely identifies exactly one record. The attributes *From_IAccess_Ref* and *To_IAccess_Ref* (numeric, 4 bytes each) are references to entries in the Internet access data model (see Section 2.1). The attributes *From_ID_Number* and *To_ID_Number* (numeric, 15 bytes each) refer to user IDs or telephone numbers of each party. For software clients and computer-to-computer calls the assigned user ID is stored. For environments where VoIP is used like traditional telephony and whenever the communication interfaces with the public telephone network the assigned telephone number is stored. The attributes *Start_Time* and *End_Time* (timestamp, 7 bytes each) refer to start and end time of the conversation.

Among the probably best known protocols for voice-over-IP are the Session Initiation Protocol (SIP, specified in [RSC+02]) and the Extensible Messaging and Presence Protocol (XMPP, specified in [SA04]). Besides these successful standardization efforts, many proprietary protocols are used, for example, Google Talk, OpenWengo, sipgate, or VoIPBuster. The popular application Skype with 245.7 million registered users in September 2007 [eBa07] uses sophisticated encryption methods [Ber05] and a complex routing algorithm for its packets, which makes it essentially impossible for an ISP to retain Skype-related data required by the EU Data Retention directive.

SIP. We mention some details for the wide-spread SIP protocol. Several header fields of SIP (defined in [RSC+02]) contain required information. From a SIP INVITE message an ISP can gather the user data for each registered user based on the IP addresses or the attributes `Via` and `INVITE` in the header. Additionally the user IDs of calling and receiving party (attributes `To` and `From`), and the start and end date of each call can be gathered by observing the transmission of INVITE and BYE messages.

4.2 Open Issues

Requirements of the EU Directive. From a technical point of view, there are some unclear or ambiguous formulations and requirements in the text of the EU directive. In **Art 5 (1) lit c Z 2 ii**, the term "Internet telephony service" is unclear. We assume that this requirement refers to the time and date of the start and end of each call. The formulation in **Art 5 (1) lit d Z 2** allows for several interpretations. We assume that it applies if a certain communication was conducted via Internet e-mail or Internet telephony.

Internet Telephony Observation. The situation with VoIP is more difficult compared to traditional telephony: Traditional telephone networks use central communication switches which may serve as observation points as the traffic is concentrated here. Popular applications like Skype, on the other hand, communicate via proprietary protocols and eventually, complicating data retention even more, decentralized or encrypted. Only if one of the conversation partners is using a traditional telephone, the observation of VoIP traffic is relatively easy.

5 Costs

We analyzed the costs incurred in terms of disk space needed for storing the data retained as well as the monetary costs involved.

5.1 Storage

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 model provider introduced in Section 1.2 and on the numbers provided in Sections 2, 3, and 4, we estimate that an Austrian ISP serving 500 000 customers with average behavior faces a permanent additional disk space requirement of about 9 gigabytes for storing Internet access related data, 624 gigabytes for Internet e-mail, and 73 gigabytes for Internet telephony in order to retain the information required by the EU directive for a six months period (including full backup).

5.2 Monetary

The costs provided in this section are divided into the following groups: Hardware (HW), software (SW), development (DEV), and general (GEN). The data in Table 3 shows that the EU Data Retention directive causes overall costs of about €972 400 in the first year, and of about €465 960 in each of the following years for the model provider. This is equivalent to costs of roughly €0.16 per subscriber per month in the first year. In terms of

personnel costs, the total salary costs for one technician were estimated as €120 per hour which corresponds to €19 200 per month per full time equivalent (FTE).

	HW	**SW**	**DEV**	**GEN**	Σ
Setup [€]	92 080	11 160	403 200	0	506 440
Operation [€/year]	120 360	0	115 200	230 400	465 960

Table 3: Total costs for storage and retrieval

The costs can be further subdivided into costs for *data storage* and costs for *data retrieval*.

Data Storage Costs. The storage process includes *gathering*, *processing*, and *archiving* of the required data. Concerning the *setup* of the data store, a storage server (amounting to €20 090, for example, product numbers XTB5220HR10A1-Z, XTB5220HR11A1SB20Z in [Sun07]), a data acquisition server (€15 990, for example, product number T20Z108B-16GA2G in [Sun07]), and various network analyzers (€40 000, extrapolated from [Ver06]) are needed. For the operation of the storage server, a database software license (€11 160, see [Ora07]) is needed. Finally, the project setup and deployment is estimated to occupy 2.5 FTEs for 6 months (€288 000).

For the *operation* of the data store, maintenance of the storage server (€110 per month, for example, product number W9D-ST5220-N-24-2G in [Sun07]) and the acquisition server (€160 per month, for example, product number W9D-T2000-8-24-3G in [Sun07]) has to be paid. Additionally, the network analyzers have to be maintained too which takes about $1/2$ FTE per month (€9 600 per month).

In summary, the data storage costs can be divided into setup and operating costs. The former amount to €375 240 and the latter to €118 440 per year.

Data Retrieval Costs. The retrieval process includes *extraction* of the requested data from the data warehouse and *delivering* it to government authorities. In [Mar06] it was estimated that one full-time employee is able to handle only two requests for retrieving data retained per day, which seems rather low. We therefore assume in this paper that 15 requests can be handled per FTE per day. One FTE per month (€19 200) was assumed for handling the requests by government authorities.

Concerning the *setup* of the data retrieval system, a data access server (amounting to €16 000, for example, product number T20Z108B-16GA2G in [Sun07]) is needed and a user-interface for data retrieval has to be developed (€115 200, 1 FTE for six months). For the *operation* of the system, maintenance of the data access server (€160 per month, see W9D-T2000-8-24-3G in [Sun07]) has to be paid. Maintenance of the user-interface requires about $1/2$ FTE per month (€9 600 per month).

In summary, the data retrieval costs can be divided into setup and operating costs. The former amounts to €131 200 and the latter to €347 520 per year.

6 Conclusions

The Data Retention directive 2006/24/EC of the European Parliament 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 controversial in many different aspects. In this paper, we focussed on the technical aspects, analyzing which data exactly has to be / can be retained by whom. On the one hand, the affected data is not always defined unambiguously. It is explicitly forbidden to store content data, but from a technical perspective the border between content and traffic or location data is sometimes blurred in modern electronic communication infrastuctures. On the other hand, ambiguities exist because the statements of the directive are in part not clearly related to underlying technical aspects.

Concerning Internet access, the implementation of the EU directive does not pose any severe technical problems. However, it seems questionable whether the basic objective can be achieved with the regulations. Persons with basic technical knowledge can gain Internet access which remains largely undetected on the basis of the data retained.

Concerning Internet e-mail, the situation is more complicated. The technical realization of the Internet e-mail system is based on a distributed approach, and there is no central authority overseeing the global traffic. Consequently, it is not always possible for a single ISP to store the full extent of the data required by the EU directive. Moreover, the protocols involved apply very tolerant regulations to the location data contained in e-mail messages. Thus, a lot of the available information may be subject to manipulation.

Internet telephony is a dynamically evolving area. Some standardization efforts exist, but many widespread applications are based on a big variety of different protocols and tools. Many of these applications are proprietary and / or encrypted. The contained information is therefore not accessible to ISPs, which makes the task of data retention imposed by the EU directive very difficult and, in some scenarios, even impossible. Traffic observation becomes particularly difficult if the communication partners do not use a traditional telephone line, but technologically more advanced setups, such as peer-to-peer approaches.

Bearing the unresolved issues discussed in this paper, the overall disk space requirements caused by an implementation of the EU directive are estimated as follows: A model ISP as introduced in Section 1.2 serving 500 000 customers with average behavior needs permanent additional disk space of about 706 gigabytes for storing data required by the EU directive over a six months period. The setup and operation of a data storage and an appropriate data retrieval system amount to additional costs of about €972 400 for the first year and about €465 960 for subsequent years.

On the basis of the data and cost model described here, various interesting questions can be addressed in the future, such as the scaling behavior of the costs in terms of the size of the ISP and in terms of the duration of the retention period, or the costs caused by a single data request from government authorities.

Acknowledgment. We thank the Internet Service Providers Austria (ISPA) for supporting this research.

References

[Ber05] T. Berson. Skype Security Evaluation, October 2005. http://www.iem.pw.edu.pl/~kozlowk3/biblioteczka/czekajace/Skype_Security_Evaluation.pdf.

[Cri03] M. Crispin. Internet Message Access Protocol. RFC 3501, March 2003. http://www.ietf.org/rfc/rfc3501.txt.

[eBa07] eBay Inc. Third Quarter 2007 Financial Results, October 2007. http://investor.ebay.com.

[Eur06] European Parliament and the Council of the European Union. Directive 2006/24/EC of the European Parliament and of the Council, March 2006. http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/l_105/l_10520060413en00540063.pdf.

[FGM+99] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol. RFC 2616, June 1999. http://www.ietf.org/rfc/rfc2616.txt.

[Hof02] P. Hoffman. SMTP Service Extension for Secure SMTP over Transport Layer Security. RFC 3207, February 2002. http://www.ietf.org/rfc/rfc3207.txt.

[Kle01] J. Klensin. Simple Mail Transfer Protocol. RFC 2821, April 2001. http://www.ietf.org/rfc/rfc2821.txt.

[Mar06] Marpij, Insight, Boivin, and Associés. Evaluation of the Economic Impacts of the Data Retention Obligations Relating to Electronic Communications, September 2006.

[Mes07] MessageLabs. Monthly Intelligence Report, July 2007. http://de.messagelabs.com/mlireport/MLI_July2007_DE.pdf.

[MR96] J. Myers and M. Rose. Post Office Protocol. RFC 1939, May 1996. http://www.ietf.org/rfc/rfc1939.txt.

[New99] C. Newman. Using TLS with IMAP, POP3 and ACAP. RFC 2595, June 1999. http://www.ietf.org/rfc/rfc2595.txt.

[Ora] 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.

[Ora07] Oracle Corporation. Oracle Technology Global Price List, August 2007. http://www.oracle.com/corporate/pricing/technology-price-list.pdf.

[Rad04] Radicati Group. Email Archiving Corporate Survey, 2004-2005, October 2004. http://www.radicati.com/uploaded_files/news/EA_SurveyPR.pdf.

[Res01] P. Resnick. Internet Message Format. RFC 2822, April 2001. http://www.ietf.org/rfc/rfc2822.txt.

[RSC+02] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, June 2002. http://www.ietf.org/rfc/rfc3261.txt.

[Run04] Rundfunk und Telekom Regulierungs-GmbH (RTR). Austrian Numbering Plan, May 2004. http://www.rtr.at/web.nsf/englisch/Telekommunikation_Nummerierung_Nationale+Nummern_nationaleRufnummern_E129.

[SA04] P. Saint-Andre. Extensible Messaging and Presence Protocol (XMPP): Core. RFC 3920, October 2004. http://www.ietf.org/rfc/rfc3920.txt.

[SM07] R. Siemborski and A. Melnikov. SMTP Service Extension for Authentication. RFC 4954, July 2007. http://www.ietf.org/rfc/rfc4954.txt.

[Sta06] Statistik Austria. IKT-Einsatz, 2006. http://www.statistik.at/dynamic/wcmsprod/idcplg?IdcService=GET_NATIVE_FILE&dID=48168&dDocName=019136.

[Sta07] Gerald Stampfel. Implementation and Consequences of the EU Directive 2006/24/EC in the Context of Internet Access & E-Mail. Master's thesis, Vienna University of Technology, 2007.

[Sun07] Sun Microsystems, Inc. Sun Enduser Price List U.S., August 2007. http://blogs.sun.com/marler/resource/price_lists/USD_MASTER-pricelist_US.pdf.

[Ver06] Verdonck, Klooster, and Associates. Study into the National Implementation of the European Data Retention Directive, October 2006.