Collaboration and the Knowledge Economy: Issues, Applications, Case Studies
Paul Cunningham and Miriam Cunningham (Eds)
IOS Press, 2008 Amsterdam
ISBN 978–1–58603–924-0

Linkage Control – Integrating the Essence of Privacy Protection into Identity Management Systems

Marit HANSEN
Unabhängiges Landeszentrum für Datenschutz (ULD) Schleswig-Holstein,
Holstenstr. 98, 24103 Kiel, Germany
Tel: +49 431 988 1214, Fax: +49 431 988 1223, Email: marit.hansen@acm.org

Abstract: In the digital world, linkage of data may pose threats to the privacy of individuals. Thus, linkage control by the individuals concerned, based on transparency of the actual and planned data processing, is the main requirement to maintain their private sphere. Today's user-centric identity management systems provide some control for users, but still lack thorough concepts of linkage control. This text introduces the phases of data processing relevant to linkage. After discussing current features of user-centric identity management concepts, an extension towards better and more comprehensive linkage control by individuals is proposed, taking into account information sources from all phases of data processing. Further, economic aspects are briefly sketched. Finally, recommendations for developers and policy makers conclude the text.

1. Introduction

The Eurobarometer survey on data protection from 2008 affirms results from other studies that a majority of EU citizens are concerned about privacy issues [1]. However, people often are not aware of actual or potential risks to their privacy, and even if this is the case, they regularly do not know how to act or react to protect themselves.

Risks to privacy usually stem from abuse of personal data, i.e., data related to individuals [2]. Data controllers often can directly link these data to their owners, namely the individuals concerned. Otherwise different data portions may be linked and accumulated into profiles that give information on the associated individuals. In many cases it is possible to identify individuals from the linked information in the profiles. Different European laws regulate the treatment of personal data. However, it is not enough to rely on European legislation in a globalised world, the more so as laws alone usually are not appropriate safeguards if not implemented in business processes and technologies.

Individuals usually have an intuitive understanding of links and linkabilities, but this understanding does not work well for the digital world with so many potential data controllers and a growing variety of identifiers and attached identity attributes, e.g., as being a citizen of a State, a customer of a company, or a user of an Internet service. User-centric identity management systems (IMS) can support users to better understand privacy risks and to act accordingly.

In the following, it is shown that even today's solutions for *privacy-enhancing* IMS have to be extended by more comprehensive possibilities of linkage controls to truly enable individuals to maintain their private sphere. In sketching this vision, we primarily address developers, vendors, and policy makers. This text is organised as follows: Section 2 introduces the concept of linkage control. Section 3 presents the relevant features in the IMS of the project PRIME – Privacy and Identity Management for Europe. Section 4 elaborates on enhancements of IMS for user-controlled linkage. Section 5 elicits relevant economic aspects. Finally, Section 6 concludes the text and gives recommendations.

2. Important Terms and Concepts

This section introduces basic terms in the field of linkage. A general model for enriching information illustrates important phases and their relation to linkage with various facets. Further it is elaborated why linkage control is the essence of privacy protection.

2.1 Terms

To link entities means to connect those entities or to establish a relationship between them. Usually **linkage** – the act of linking – is done for a specific purpose, and this purpose determines which entities can or will be linked. For example a car can be linked to its owner by checking out the registration number at the authority which stores that information; the IP address can be linked to a computer that has been assigned that number by an Internet Service Provider; transactions of the same eBay user can be linked and compiled into a profile of that user. A typical way of linkage is to relate different portions of data which have the same identifier, but it is also possible to establish links because of other information, e.g., concerning time or location.

Note that an established link does not mean that the linked portions of data belong together. For example two data sets with information on a "John Smith" may be linked even if there are two persons named "John Smith", not knowing each other, totally unrelated except for accidentally having the same name.

Linkability (i.e., the possibility to link) and its negation **unlinkability** are dependent on the attacker's perspective, i.e., the data which are available for him and further knowledge on ways of successfully linking those data [3].

Comparing linkage of data with linkage of chain links, it seems logical that there can be also some "**de-linkage**", i.e., removing a once established link. This can be achieved by separating data into different databases that cannot be accessed by the same person or by deleting components of a data profile. If separation or deletion is not possible, the validity of the linkage may be challenged by providing contradicting information, e.g., by injecting disinformation. De-linkage often cannot be guaranteed because the linked information may already be memorised by people or copied and further processed by ICT (information and communication technology) systems.

To complete the terms, the uncommonly used word "**de-linkability**" stands for the possibility of de-linkage while "**un-de-linkability**" means the opposite [4].

Linkage control means to know about linkages performed or planned, to influence its conditions (at least in a defined and known scope) and to be able to check afterwards whether the linkage was done properly and as agreed upon. A necessary requirement for linkage control is transparency, i.e., clarity on the terms of data processing. The most effective way of preventing linkage is to prevent **linkability**, i.e., the possibilities of linkages. Related is also **observability control** because if no information is observable, there will be nothing to link. Linkage control can also mean having guarantees that some data portions actually are or will be linked. This may be the case for reputation systems where an individual should not be able to remove unwelcome entries.

2.2 A General Model for Enriching Information Put into the Linkage Context

Digital identities that represent people in the digital world are often linked with information about this very same person, e.g., social contacts or actions performed under that digital identity. In addition, this information can be further specified or extended by linking it with other data sources, e.g., other digital identities of the same person, and utilising scoring models or other sophisticated algorithms which analyse the data.

Figure 1 shows the typical data flow when enriching information for the purpose of generating decisions, as this is done multiple times a day in common data processing systems. Not always the phases are as pronounced as in profiling and scoring systems:

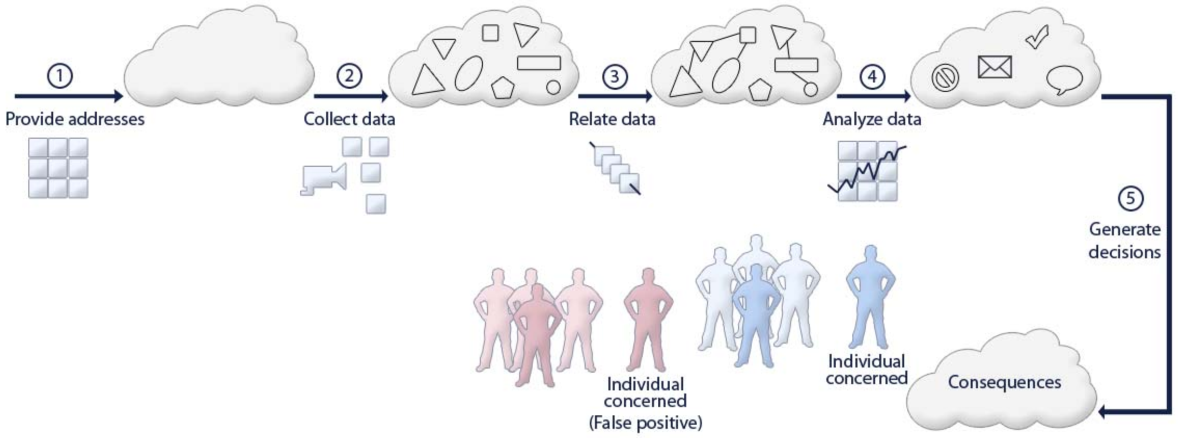

Figure 1: Model of information flow relating to linkage

The successive phases of this model are explained in Table 1, which also discusses the relation of each phase to linkage.

For discussing linkage properties and objectives, it is important to make clearly visible who can access which data and perform which actions on these data. In all identified steps in the model workflow presented in Figure 1, the actors contribute to some aspect of linkage and may have to be cautious to avoid undesired effects. In case of a mistake, it may be hard for individuals concerned to find the error and its cause in this workflow and to achieve that appropriate corrective measures are being taken.

Table 1: Phases in the model of information flow

	Description	**Relation to linkage**
Provide addresses ⊖	In ICT systems, each object has addresses which acts as distinguishing information in a specific scope. Addresses can be names or identifiers which may enable referencing, distinguishing and identifying objects. Often the addresses are assigned according to a defined address schema, e.g., IP addresses or e-mail addresses. Often the addresses contain more information than it is necessary for the purpose of distinguishing.	a) The address may represent an individual, an action of an individual, an object possessed by an individual. b) Information on the assignment of addresses contains the link between the address and what it stands for. c) Even without an assignment table links can be established, e.g., between data sets when addresses occur repeatedly. d) Already the address schema and the assignment process set the context for possible linkages and the interpretation of those links, e.g., whether addresses are unique for single individuals or whether addresses can be taken by other individuals (voluntarily or not, as in the case of identity theft).

❷ Collect data	In addition to the data processing performed on legal grounds or with the user’s consent there may be hidden collection of data. There is probably no single entity which can monitor all user actions or data transfers, but there are many which can observe some parts of users’ lives or ICT systems. Users are often unaware of data trails they are leaving, e.g., when browsing the Internet or when using mobile phones. Further many people provide personal information to social networks or blogs where they are publicly accessible.	a) All information which is observable can be monitored and collected. b) The data collector is not necessarily related to the party assigning addresses or defining address schemas. c) Once data are disclosed, it usually cannot be guaranteed that they won’t be part of some data collection and transfer – either by professional data controllers or by other individuals. d) Also typically temporary data can be stored permanently. e) The information whose personal data are collected may be available in this stage as well. Otherwise information such as identifiers, location, time etc. can be gathered and stored together with the observed data. Then the relation to a person may become clear after linking the data in phase 3.
❸ Relate data	The raw data being collected in the previous phase are linked in this phase. Usually this process of relating data is done for a specific purpose which determines conditions for the linkage, e.g., which data are relevant, when should data be related (e.g., sameness of addresses, similarity of time information etc.). As relating data in a correct way (determined by the purpose of data processing) depends on specified conditions and assumptions, specific linking algorithms can be used.	a) Relating data is linkage on the data set level. b) The related data can be directly linkable to an individual or they can be a pseudonymous profile of an individual with no knowledge of the identity behind. c) The process of relating the data may be conducted by special parties. d) Further the algorithms to be used may be provided by yet other parties which may not be related to the other phases of data processing.
❹ Analyse data	Also the analysis of the compiled profiles and other data portions usually is driven by the given purpose. Different methods can be employed such as scoring functions, expert systems or neural networks, possibly being providers by specific parties. In this phase decisions are prepared.	a) As the data analysis bases on the linked data from the previous phase, the linkage properties are inherited. b) In addition the analysis algorithms may link the data to other information, e.g., facts or assumptions from other sources or rules being applied to the data. c) The process of analysing data may be conducted by special parties. d) Further the algorithms to be used may be provided by yet other parties which may not be related to the other phases of data processing.
❺ Generate decisions	Actual decisions are generated in this phase. They may affect single or multiple individuals. The decisions are not necessarily fair, and sometimes they do not base on accurate data or on correct assumptions when relating or analysing data. In these cases individuals being “false positives” may suffer from consequences of the generated decisions, e.g., if they have to pay more than others for the same service or if they won’t get a service at all.	a) The decision maker decides on basis of the information available at that stage, being provided by previous phases. b) This phase links the data processing to real consequences which may affect individuals also in their real lives, not only their digital representations. c) By the decision, the link is (re-)established to the individuals concerned, whether all phases were performed correctly on accurate data or not.

2.3 Linkage Control as Essence of Privacy Protection

There are various definitions for the right to privacy, each of them focusing on specific aspects. Two of the most mentioned concepts are firstly the “right to be let alone” [5] and secondly the right “to control, edit, manage, and delete information about them[selves] and decide when, how, and to what extent that information is communicated to others” [6] or similarly the “right to informational self-determination” stemming from the 1983 ruling of the German Federal Constitutional Court, demanding that each person can at any time ascertain who knows what about him or her. Both definitions are related to linkage control:

For the “right to be let alone”, individuals should be able to control (or prevent) the linkage at least in the last stage (phase 5 in Figure 1) when decisions may concern them. For the informational self-determination, in particular earlier phases are relevant, too, because all available data (e.g., disclosed information on the respective individual) is the material for the knowledge acquired by other parties.

Also extended perspectives such as the privacy categories from [7] are based on linkage: The supplementary aspect of group profiling and social sorting describes possibly anonymous profiles which may contain information that can be used to discriminate against specific individuals. Or the link to the individual can be established later, e.g., by additional algorithms, computing power or data. Here all phases from Figure 1 are relevant. Further the need for defining what is public and private [7] again is based on linkage control so that individuals involved in participatory processes can prevent unfavourable consequences.

3. Linkage Control Features in PRIME’s Identity Management System

In the digital world, linkage control for users is currently much more difficult due to massive data processing which is mainly opaque for them. Privacy-enhancing identity management systems strive for linkage control by the user. For this reason we list in this section a few interesting functions which support linkage control.

Identity management means managing various partial identities (usually denoted by pseudonyms) of an individual, i.e., administration of identity attributes including the development and choice of the partial identity and pseudonym to be (re-)used in a specific context or role [3]. Identity management systems can be distinguished by the degree of control from individuals respectively organisations when administrating the partial identities. It always depends on the context as well as on the perspectives of the parties involved how much linkage, linkability or unlinkability is desired in a specific situation.

Meanwhile so-called “user-centric identity management systems” [8] dominate the landscape where users are given at least some control on their identity data. Typically users can decide on how much information they are willing to disclose in a specific situation. However, even big systems such as Microsoft’s CardSpace by now do not support users in interpreting the privacy policy on the service’s side with its linkage-related information.

An enhanced approach is shown in the project “PRIME – Privacy and Identity Management for Europe” [9], as depicted in PRIME’s White Paper [10], based on some fundamental work from David Chaum since the early 1980s [11]:

- To achieve better linkage control, the workflows within organisations can be organised in a way that they prevent globally unique identifiers, but instead restrict the identifiers’ scope to the necessary domain. In different contexts, different pseudonyms can be used against unwanted context-spanning linkages (cf. the “Identity Protector” [12]).
- Undesired linkage is also prevented by private credentials, i.e., certificates proving identity claims (e.g., “being of age”) without revealing information that may identify the individual [11], [13]. Multiple private credentials can be created from a single master certificate that are neither linkable to each other nor to the issuance interaction of the master certificate.
- Individuals are supported in knowing beforehand the conditions of data processing by privacy policies that are both understandable by human beings [14] and machine-readable, i.e., they can be interpreted by the user’s system. In addition, the organisation can automatically enforce its privacy policy – and all included statements concerning limitations of linkages – with appropriate tools.
- Such privacy policies can be cryptographically “stuck” to data sets [15], [16]. Thereby these sticky policies can travel together with the data they apply to.

- PRIME integrates multiple transparency functionalities [17] which make users understand better possible linkages and to react accordingly: The main tool is the so-called “Data Track”, a logfile of prior transactions which stores a record of what identity information has been disclosed to whom and under which conditions. This logfile enables users to review later what they consented to. It also can warn users against too much data disclosure. Other transparency tools conceptualised in PRIME are the “Security Feed” based on RSS which gives machine-readable information on privacy and security incidents, or the support of individuals to exercise their privacy rights, i.e., managing requests towards the data controller demanding access to their personal data, rectification or erasure as well as giving and withdrawing consent. This kind of tools will be further developed in the FP7 project PrimeLife [18].

4. Extension of IMS to Enable Linkage Control by the User

How can the described features in Section 3 be mapped to the linkage model in Section 2, which illustrates typical data processing phases in the world? Although PRIME’s features seem to already cover a lot of what is necessary to enable true linkage control, they cannot give a complete picture, as they mainly address the direct relationship between user and service. An exception is the “Security Feed” which may provide information also from other sources than the service itself. The possible (and on a large scale meanwhile common) data processing by others, be it secret services of other nations or curious peers, is not shown to the user in current IMS. Today, data transfers and integration of other parties are at best roughly described in privacy policies, and information on analysis algorithms actually being applied or assumptions made is usually missing at all.

For an integral whole of linkage control by users, several extensions should be considered that address not only IMS software designers, but also legislators, policy makers, data controllers, standardisation bodies, and data protection authorities:

1. **Transparency on linkability and linkage**
   - Information on possible and actual linkages as well as de-linking options should be available by the individuals concerned. In an abstract way (without mentioning personal data of individual cases) this could be provided in public databases. At least concerning governmental activities this should be legally demanded. Information on quantification of linkability [19] would be helpful.
   - Privacy breaches should be communicated to the individuals concerned.
   - IMS should be able to inform users about possible privacy risks by interpreting the information sources mentioned in the bullet points before.
   - Data controllers should always document the sources of their data and algorithms used as well as the actual recipients (not only categories of recipients as currently demanded by privacy law). They should be able to prove that their data processing is lawful. In case of questions by users or supervisory authorities it should be possible with little effort to review the full audit trail covering all phases of data processing (cf. Figure 1) until decisions are generated.
   - The information obligations should not be limited to clearly and directly personal data, but should also comprise other data suitable to affect individuals.
   - Even if no privacy or security risks occur, individuals should be informed on data processing. This could be implemented, e.g., as an “itemised statement” – similar to telecommunication bills – sent by data controllers to users, stating who has accessed the individual’s personal data for which purpose and giving further information.
   - For enabling the IMS to orchestrate the available information on processes and actual and possible linkages, standardised formats and ontologies will be required.

2. **Control of linkage**
   - Data controllers as well as standardisation bodies should take care of observability and linkability issues already when defining address schemas, processes and protocols. In relevant areas privacy impact assessments should be conducted.
   - Cross-jurisdiction data processing transferring data out of the area where users can exercise their linkage control and where privacy regulations can be enforced should be avoided. For instance, the European Union should offer their citizens ways to keep their personal data in the EU jurisdiction – also when using mighty search engines, transferring money within the EU or booking inner-European flights.
   - Users should be informed on how to check all data processing concerning them and be provided with effective possibilities for correcting occurred errors and for redress. Processes for checking and redress should be handled in an easy way.
   - IMS should raise the users' awareness and support them in their linkage control. Good usability as well as high data security and reliability are prerequisites.

5. Economic Aspects of Linkage Control

Today, we are far from offering users linkage control regarding their privacy – neither in the traditional nor in the ICT-enhanced world. The vision of full linkage control bases on manageable ICT systems both at the service's and the user's sides, and it requires a plurality of available information channels which can be interpreted by the user's IMS. On the one hand this calls for new kinds of information providers as well as providers of the supporting infrastructure. These providers have to develop workable business models – supply and demand here is still unknown territory. But linkage control does not only address new services: Starting from existing systems, each business process or ICT system designer should consider supporting linkage control from the outset through the different phases of information flow.

Currently there is no comprehensive analysis of economic aspects regarding linkage control. However, a few specific features are being explored. For instance, transparency of security risks is being discussed for several years – also from the economic point of view [20] – and has gained momentum in the process of reviewing the ePrivacy Directive where data breach notification provisions are being debated. Further, the discussion of business models of privacy-enhancing technologies addresses many linkage control issues [21]. Clearly, the carrot-and-stick approach is promising, i.e., a mixture of firstly economic incentives and secondly sanctions when not adhering to the law or not meeting the required state-of-the-art. This demands a consensus among policy makers on the value of privacy and the consequence of exercising linkage control.

6. Conclusions and Recommendations

Our information society with its data processing is to a great extent based on linkage. Linkage control is one of the most important concepts for self-determination of individuals. In the digital world full of identifiers for digital identities which often can easily be linked, better linkage control by individuals is crucial for maintenance of their private sphere. Control is based on transparency and checkability. This requires that the complex world of today's data processing with manifold actors has to provide all relevant information to check correctness and fairness of decisions.

Privacy-Enhancing Identity Management Systems could act as the users' assistants and guardians if being enhanced to interpret sources with all information relevant to them and supporting them in exercising their control. In a way this would mean to mimic "world knowledge" as linkage aspects cover so many facets of data processing. However, this vision is not totally unrealistic in a world of semantic web, ontologies being standardised and governmental processes in the scope of the EU Services Directive being translated into XML. True linkage control needs information, much more than it is available today. Here a balance between possible trade secrets from data controllers when applying scoring algorithms and the requirement of checkability by supervisory authorities and individuals concerned has to be found. Also a societal consensus on possible limitations of linkage control by individuals should be achieved.

Policy makers as well as system designers should appreciate the value of linkage control and pick up the concept in their respective areas. More work is needed when it comes to linkage-relevant interaction between peers instead of data controllers because the former are typically not subject to privacy regulation.

References

[1] Eurobarometer, Data Protection in the European Union – Citizens' Perceptions. Analytical Report, Flash Eurobarometer No. 225, Survey conducted by the Gallup Organization Hungary upon the request of DG Justice, Freedom and Security, Feb. 2008, http://ec.europa.eu/public_opinion/flash/fl_225_en.pdf.

[2] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data. 01248/07/EN WP 136, June 20, 2007, http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf.

[3] A. Pfitzmann, M. Hansen, Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology. Working Paper v0.31, February 15, 2008 (first version from 2000), http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

[4] M. Hansen, S. Meissner (Eds.), Verkettung digitaler Identitäten. Report commissioned by the German Federal Ministry of Education and Research, October 2007, https://www.datenschutzzentrum.de/projekte/verkettung/2007-uld-tud-verkettung-digitaler-identitaeten-bmbf.pdf, Lulu Inc., Feb. 2008, ISBN 3000234063.

[5] S. Warren, L. Brandeis, The Right to Privacy. In: Harvard Law Review, Vol. 4, 1890, pp. 193-220.

[6] A.F. Westin, Privacy and Freedom. Atheneum, New York, 1967.

[7] D.J. Phillips, Privacy policy and PETs – The influence of policy regimes on the development and social implications of privacy enhancing technologies. In: New Media & Society, Vol. 6, No. 6, SAGE Publications, London, Thousand Oaks, CA and New Delhi, 2004, pp. 691-706.

[8] A. Jøsang, S. Pope, User Centric Identity Management. In: Proceedings of AusCERT Conference 2005, Brisbane, Australia, May 2005.

[9] PRIME – Privacy and Identity Management for Europe, FP6 IST project, https://www.prime-project.eu/.

[10] R. Leenes, J. Schallaböck, M. Hansen (Eds.), PRIME White Paper V3. May 2008, https://www.prime-project.eu/prime_products/whitepaper/.

[11] D. Chaum, Security Without Identification: Transaction Systems to Make Big Brother Obsolete. In: Communications of the ACM, Vol. 28, No. 10, Oct. 1985, pp. 1030-1044.

[12] H. van Rossum, H. Gardeniers, J.J. Borking et al., Privacy-Enhancing Technologies: The Path to Anonymity. Volume I & II, Achtergrondstudies en Verkenningen 5b, Registratiekamer, The Netherlands & Information and Privacy Commissioner/Ontario, Canada, Aug. 1995.

[13] J. Camenisch, A. Lysyanskaya, Efficient non-transferable anonymous multi-show credential system with optional anonymity revocation. Research Report RZ 3295 (# 93341), IBM Research, Nov. 2000.

[14] Article 29 Data Protection Working Party, Opinion on more harmonised information provisions. 11987/04/EN WP 100, Nov. 25, 2004, http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2004/wp100_en.pdf.

[15] G. Karjoth, M. Schunter, M. Waidner, Platform for Enterprise Privacy Practices: Privacy-enabled Management of Customer Data. In: Proceedings of 2nd Workshop on Privacy Enhancing Technologies (PET 2002), LNCS 2482, Springer, pp. 69-84.

[16] M. Casassa Mont, S. Pearson, P. Bramhall, Towards Accountable Management of Identity and Privacy: Sticky Policies and Enforceable Tracing Services. HPL-2003-49, Trusted Systems Laboratory, HP Laboratories Bristol, 2003, http://www.hpl.hp.com/techreports/2003/HPL-2003-49.pdf.

[17] M. Hansen, Marrying Transparency Tools With User-Controlled Identity Management. In: Proceedings of the Third IFIP WG 9.2, 9.6/11.6, 11.7/FIDIS International Summer School, Aug. 2007, IFIP International Federation for Information Processing, Vol. 262, Springer; 2008, pp. 199-220.

[18] PrimeLife – Privacy and Identity Management in Europe for Life, FP7 IST project, http://www.primelife.eu/.

[19] S. Berthold, S. Clauß, Linkability Estimation Between Subjects and Message Contents Using Formal Concepts. In: Proceedings of the 2007 ACM Workshop on Digital Identity Management, 2007, pp. 36-45.

[20] R. Anderson, R. Böhme, R. Clayton, T. Moore, Security Economics and the Internal Market. Report for ENISA, Feb. 2008, http://www.enisa.europa.eu/doc/pdf/report_sec_econ_&_int_mark_20080131.pdf.

[21] P. Ribbers (Ed.), Business Processes and Business Case. PRIME Deliverable, May 2008, https://www.prime-project.eu/prime_products/reports/reqs/pub_del_D2.2.a_ec_WP2.2_v5_Final.pdf.