Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

PETweb II deliverable D8: Case study — Privacy-relevant information flow in identity management systems

Note no	DART/06/10
Author	Bjarte M. Østvold
Date	October 1, 2010

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in the areas of information and communication technology and applied statistical modelling. The clients are a broad range of industrial, commercial and public service organizations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have for us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**PETweb II deliverable D8: Case study — Privacy-relevant information flow in identity management systems**
Author	**Bjarte M. Østvold**
Quality assurance	Lothar Fritsch
Date	October 1, 2010
Publication number	DART/06/10

Abstract

This document is deliverable D8 in the PETweb II project, financed by the Research Council of Norway.

Keywords	identity management, privacy, information flow, privacy-enhancing technology
Target group	All
Availability	Open
Project	PETweb II
Project number	320426
Research field	information security and privacy
Number of pages	15
© Copyright	Norwegian Computing Center

Contents

1 Introduction

How information flow underpins most privacy and security concerns in systems. If all information is static, no one may learn anything that they didn't know already[1] Thus, this document considers what information flow means in identity management, and looks briefly at three case studies in order to extract principles for the design of privacy-aware identity management system from the viewpoint of flows.

2 Terminology

2.1 Information, data, and interpretation

Information equals data plus interpretation. This equation is the core tenet of data analysis: we must interpret data in order for it to become information. In general, *interpretation* is about assigning meaning to data. In this report we assume that meaning is at least partly sensitive, for example, the meaning could include the name of a person or a partial record of that person's activities.

Note that when data can be trivially interpreted, for example, data in the form of a document, data too can be sensitive. When the interpretation of data is non-trivial, for example, requiring a cryptographic key, then the data is not sensitive while the information may still be sensitive.

2.2 Identities and identifiers

An *entity* is a person, an organisation, a company, a group of people, or a grouping of other entities. Associated with an entity there may be information that the entity wants to keep secret, sometimes called *private information*.

For each entity there is a unique piece of information, the *identity*, that unambiguously determines the entity. An *identifier* is piece of data that may trivially be interpreted as the identity of an entity. A *pseudonym* is an identifier that only determines an entity non-trivially, for example, using decryption. Typically identifiers are syntactic object that follow a described format. Fig. 1 depicts these terms.

[1] Except for the meta-knowledge that information stopped flowing.

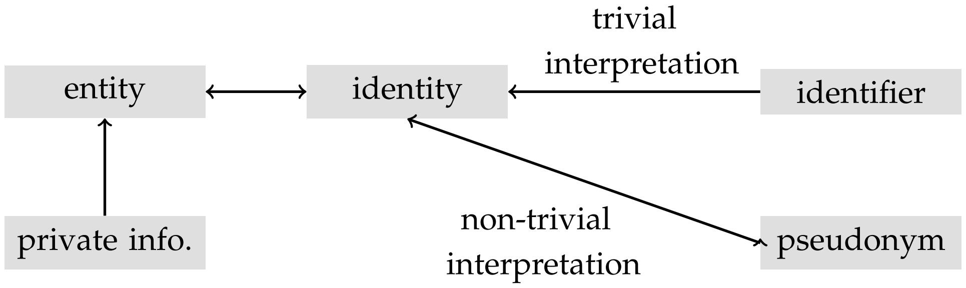

Figure 1. Terminology related to identity.

2.3 Identity management

In this report we take a technical viewpoint on identity management, focusing on security and privacy. We believe, however, that out terminology is useful also in the broader organisational and legal context. *Identity management (IdM)* is about manging the access of identities, that is, the access of entities, to the resources of some system. It is useful to consider such a system as consisting of two parts or subsystems: The identity management subsystem is the part working with identities whereas the domain subsystem or domain service is the part that delivers some service to entities, a service that involves resources that are not only identity-related, but belong to some domain, for example, health care. Here, our interest will lie with the IdM part of the system, and we refer simply to this part as the *IdM system*, or simply the system where there can be no confusion.

In general the information processed by an IdM system will be, at least in part, sensitive. Examples of sensitive information are the identities of entities, *attributes* that the IdM system associates with identities, and the resources the system associates with entities. Examples of sensitive information are access times, cryptographic keys, names, or other personal information. In addition, the system may build sensitive data as part of its operation, for example, logs of system operations.

2.4 Information flow

In computer science the classical meaning of flow analysis is to analyse how data flows inside a program, for example, to determine what variables are live at certain points in the program (Nielson et al., 2005). Here a live variable is a variable that may be needed in the future, that is, needed at program points reachable from the current point. Many kinds of flow analyses are used in compiler optimisation. In computer security, the term information flow analysis refers to the analysis of program with variables classified into different security levels, and the goal of analysis is to ensure that information does not flow from, for example, low-security variables to high-security ones (Sabelfeld and Myers, 2003).

One could imagine *privacy-relevant information flow* being the case where there are variables corresponding to each entity and that the information in these variables is not allowed to flow to other entities, but only to variables controlled by the IdM system itself. This view is not entirely satisfactory since it does not prevent the system for accumulating unnecessary privacy-relevant information—a behaviour that is risky, but does not explicitly disclose the data. More on this in Section 4.

3 Example scenarios

3.1 Government web portals: ID-porten

The term *ID-porten* stands for all the public sector activities in identity management and citizens' use of electronic identities (eID). The government has chosen to base this initiative on federated identities[2], single-sign on, and single log-out off for public sector services. Several existing eIDs may be used to sign on to ID-porten, and public service providers may offer personalised services, for example, tied to personal heath information.

ID-porten allows citizens to login using identities supplied by accredited third parties, and it then translates these identities into ID-porten identities. Currently Buypass and Commfides have been accredited as such 'identity suppliers' to ID-porten. ID-porten also lets citizens login directly using its own identity scheme, called MinID. The *raison d'être* of ID-porten is to allows citizens to log into and use certain government services, like the Norwegian Tax Administration. An overview appears in Fig. 2.[3] In the figure, the regular lines indicate legitimate flows of information while the dotted lines indicate suspect flows.

The design of ID-porten is based on the OASIS SAML 2.0 standard for web services. The standard defines a way to describe and exchange security-related information between services, based on XML. The key services of ID-porten itself are:

- Authentication/login based on attributes.
- Signing of documents and signature verification.
- Facilitating encryption and decryption.

The SAML standard includes a central service, the identity provider (IDP) that

[2] Federated identity means connecting identities across services and realising that they are the same.

[3] Fig. 2 is a minor adaptation of a figure made by Lothar Fritsch.

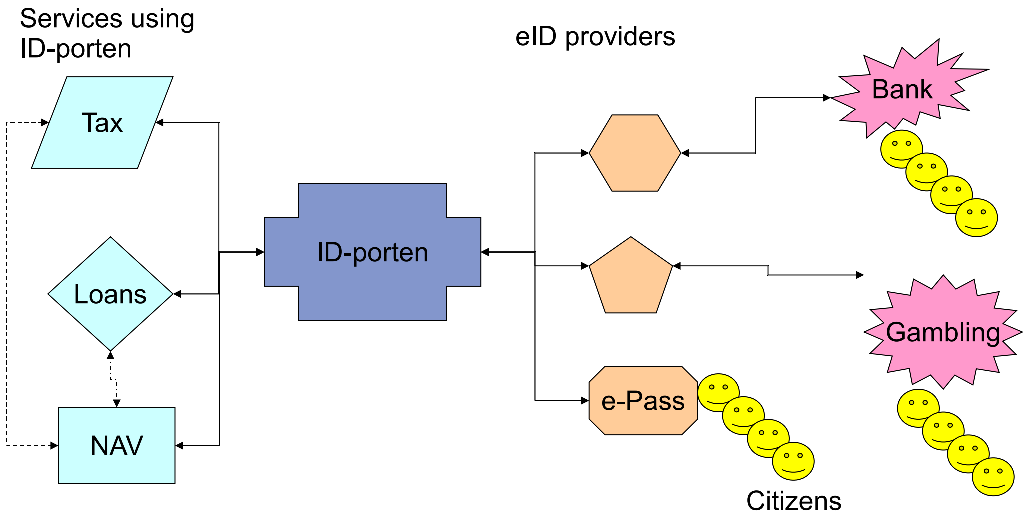

Figure 2. Overview of ID-porten and associated services.

each service provider (SP) talks to. SPs need not trust each other, but each SP must trust the IDP.

The implementation of ID-porten is based the OpeanAM access management product from ForgeRock AS, the OpenDS Directory Server, and more.

3.1.1 Scenarios

The follow scenarios of ID-porten includes both the users (entities) and the administrators of the IdM system.

Single sign-on A citizen signs on using one of the authorised eIDs, then starts using one service, for example, to change her "skattekort" with the Norwegian Tax Administration (Norw. Skatteetaten). Then she decides to change her appointment with her physician.

Performance analysis The administrators at ID-porten want to improve the performance of their portal. They decide to gather some statistics about what and how much data flows between the different service providers.

3.2 The Internet of things: Electronic tickets

There is currently lots of industrial interest in radio-based tagging technology (RFID, NFC) and embedding intelligence into tags mounted on various kinds of commercial goods, from soft drink bottles, to access cards, to clothes. The network and services associated with these tags is sometimes called the Internet of things. The deployers of the technology hope to make their systems more efficient (less costly) through attaching tags to objects handled by the systems or by making people carry tagged tokens. The flow of objects or people (via tokens) through the system can then be tracked using tag readers and stored for various purposes, for example, billing, addressing, origin tracking, flow optimisation, and control.

Some tasks, for example, billing or flow optimisation, regard the system as a whole. Information pertaining to such tasks must be collected globally, that is, system wide, and analysed together. Other tasks, such as determining the origin or transport history of a package, regard only a single entity of the system and the path the entity has travelled. Information relevant to such tasks could be served by system-wide collection and analysis, but may equally well be served by local collection and analysis, with less infrastructure and greater flexibility.

The Internet of things creates new challenges for identity management:

- Tags, or sets of tags, implicitly create association to persons, either anonymous or identified.
- These associations, and how they develop over time, can reduce the anonymity of those persons.
- Current tags have little or no protection against reading their information content. This design choice helps keep down the cost of tag hardware, a key concern is mass production of tags.

In summary managing tag means indirectly managing identities, and the current state of affairs is unsatisfactory from a security and privacy viewpoint.

3.2.1 Scenarios

Transportation companies, including the Norwegian company Ruter AS in Oslo and Akershus, sell or plan to sell electronic tickets, for example, in the form of tagged paper clips or tagged plastic cards. There are two main kinds of tickets: Anonymous tickets for limited use, for example, travel for a fixed number of trips; and personal subscription tickets, for example, subscriptions for a month of service valid for one identified person.

Anonymous electronic tickets for limited use A person travels the same journey (back and forth) at about the same time of day, nearly every working day for 6 months. Furthermore, we assume that the person has a public home and work addresses.

Profiling to planning transport resources The transport organisation wants to know more about travel patterns in their network in order to plan their resource usage better. They decide to monitor all passengers using subscription cards for one month on travel.

3.3 E-health: Tracking demented patients

With the ageing population in Western society, the number of people summering from dementia is increasing. Demented patients have reduced ability to orient themselves and thus are at a greater risk to get lost, leading to risk of harm. Also, caretakers of demented patients spend resources on searching for lost patients.

New technology, combining satellites (GPS) and mobile phone in one small portable device, presents a low-cost way of tracking and finding lost patients. Such use presents a lot of ethical and legal problems, in addition to identity management problems.

3.3.1 Scenarios

Nursing homes with many demented patients are a potential user of patient tracking technology. The nursing home staff, however, are not technology experts, and we therefore assume that patient tracking will be sold as a service to the homes, with tracking data being collected and analysed, in part, at a separate site operated by the provider of the tracing service.

Register new patient A nurse registers a new patient and associates the patient's name with the tracking identifier. The nurse then takes the patient on a walk around the home premises, and later checks to see that the movement was registered correctly by the system.

Nurse scheduling As part of the work force scheduling at the home, the chief nurse extracts a list of all patient movements the last month and uses it to find periods requiring more staff or fewer staff.

Surveillance Under certain conditions designated persons or groups of persons should have access to the position of patients. Examples may include security guards investigating incidents, family members looking for a relative in a home-case situation, or rescue workers operating under emergency procedures.

4 On analysing information flow scenarios

Here, we do not analyse the concrete scenarios further, but instead we list concerns involved in modelling and analysing information flow in identity management systems, generalising concerns from the scenarios

In an identity management system, the confidential data includes, of course, identities and their associated entities. A system processing confidential data has as a central concern that it should not release such information to unauthorised entities. In particular, this means that the system, as part of its operation, should also not increase the risk of such release, and it should not increase the impact of such release if it should happen. An example of a risky system is one that accumulates logs of confidential data over a long time without the need for such storage.

Confidentiality by association risk Data associated with an identity or a pseudonym may also become confidential, since that data may either reveal the entity behind an identifier or may reveal private information about the entity.

Data aggregation risk As the IdM system operates it might increase the amount of data associated with identities, thus increasing the risk to identities.

Application-side duplication risk Applications using the IdM as part of service delivery, may be tempted to duplicate parts of the IdM system in order to realise identity-related functions that the IdM system itself does not allow.

Below we list some principles for privacy-aware identity management design with respect to information flows. When analysing scenarios, these principles are issue one should check for.

Adding an identity When a new identity is added to the system, data associated with pre-existing and unrelated identities should not change.

Deleting an identity When an identity is deleted, data associated exclusively with the identity should normally be deleted too.

Modifying data associated with an identity When data associated with an identity changes, data associated with unrelated identities should not changes.

Aggregating information IdM systems might associate information with identities, information that is not part of IdM services to applications or users. Examples of such information is logging. Such information should be kept at a minimum, and only be created as a result of contractual obligations, for example, non-repudiation, or documented operational goals, for example, caching. Convenience of implementation or "storing everything in perpetuity" are not valid arguments.

5 Outlook on automation

In this section we consider the various kinds of theory and technology available today for automating information flow analysis. A prerequisite for using such tools is to create a model of the system, abstracting away from irrelevant aspects and implementation details. That model can then be input to the types of tools discussed below.

5.1 Protocol analysis tools

One could see the IdM system and its users as agents acting out some protocol that must satisfy certain security properties. To investigate alternative *designs* for the system means to investigate various protocols that satisfy the properties.

This perspective allows to bring to bear tools developed for protocol verification, for example, a recent survey paper (Cremers et al., 2009) lists the tools Avispa (multiple variants), Casper/FDR, and Scyther. Another tools is ASTRAL Coen-Porisini et al. (1997).

In addition, the system should not leak confidential data through it operation. From this perspective, Leaks can be seen as similar to side-channels in cryptography.

5.2 Petri net tools

Petri nets seem a natural fit for modelling information flow, at least generalised variants such as coloured Petri nets where the tokens can have values in order to model the flows. CPNTOOLS is an example of a tool that allows modelling with such nets[4].

5.3 Static analysis tools

Static analysis for programs, including data flow analysis, is complex and needs to be tailored to the programming language in question. It seems therefore more realistic to use an existing language where different static analyses have already been developed. The experimental programming language Jif extends Java with types for information flow and access control[5], and it is such a language.

[4] `http://wiki.daimi.au.dk/cpntools/`

[5] `http://www.cs.cornell.edu/jif/`

References

Coen-Porisini, A., Ghezzi, C., and Kemmerer, R. A. (1997). Specification of real-time systems using ASTRAL. *IEEE Trans. Software Eng.*, 23(9):572–598.

Cremers, C. J. F., Lafourcade, P., and Nadeau, P. (2009). Comparing state spaces in automatic security protocol analysis. In Cortier, V., Kirchner, C., Okada, M., and Sakurada, H., editors, *Formal to Practical Security*, volume 5458 of *Lecture Notes in Computer Science*, pages 70–94. Springer.

Nielson, F., Nielson, H. R., and Hankin, C. (2005). *Princples of Program Analysis*. Springer.

Sabelfeld, A. and Myers, A. C. (2003). Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19.