Knowledge discovery in a circle of trust

L. Peyton & J. Hu
School of Information Technology and Engineering,
University of Ottawa, Canada

Abstract

There are now many services for individuals and businesses available over the Internet. In providing these services, personal data is exchanged. There are security and privacy concerns about protecting identity and controlling the use of personal data. The Liberty Alliance project has developed a set of standards and architecture for federated identity management. A circle of trust (CoT) is a network based on the Liberty Alliance architecture in which businesses collaborate to provide services in a manner that protects identity and carefully controls the sharing of personal data. We analyse the requirements of each stage of a knowledge discovery process in the context of a CoT, and recognize data collection as an important step with respect to protecting privacy and identity. An eHealth scenario in which data mining is used to detect prescription misuse illustrates how to implement a trusted data collection architecture in a CoT .
Keywords: circle of trust, data collection, privacy, data mining, architecture.

1 Introduction

With the growth of on-line services data is often shared and collected across organizations in distributed business to business (B2B) networks across the Internet. Data mining is an important tool for monitoring the services provided within such networks, but the data needed for data mining may come from several organizations. Because each organization operates independently, they each have only a partial view of the data involved. Data sharing is required in order to collect and analyse data. However, there are security and privacy concerns about protecting identity and controlling the use of personal data. In response to this concern, governments have enacted privacy laws such as the European Union Prime Directive on Privacy [1], the Health Insurance Portability and Privacy Act [2] in the United States and the Privacy Act and the Personal Information Protection and Electronic Documents Act (PIPEDA) [3] in Canada.

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070221

There have also been industry initiatives to control data sharing. Federated identity management can enable users and service providers to securely and systematically manage identities and personal data in a single sign on framework that controls access to personal information. The Liberty Alliance project [4] was established in 2001 to develop an open standard and set of specifications for federated identity management. A key concept in the Liberty Alliance project is a "Circle of Trust" (CoT), in which federated identity management is used to create a business to business (B2B) network of cooperating enterprises that provide integrated services to users. These cooperating enterprises have trust relationships and operational agreements established amongst them.

In this paper, we use the example of an on-line prescription scenario from the literature [5] to examine how a knowledge discovery process can proceed within the context of a B2B network. In particular, we:

1. Identify the requirements of each stage of a knowledge discovery process based on the CRISP-DM model [6] in the context of a B2B network created as a CoT using the Liberty Alliance framework
2. Propose an additional stage in the process, data collection, to address the collection of data from different organizations in a B2B network.
3. Define a "trusted" data collection architecture within a CoT, based on the creation of "Data Collector" and "Source Directory" services and a "Client Master Index" to link federated pseudonyms.

2 Background

Shearer [6] introduced the CRoss-Industry Standard Process for Data Mining (CRISP-DM) that organizes the data mining process into six stages: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Zaidi et al [7] proposed a distributed agent infrastructure for data mining to provide healthcare-oriented decision-support services, but privacy issues were not addressed. One approach to address privacy, based on statistical methods, is to alter the data before it is sent to the data mining tool, so that real values are protected, but important statistics are still preserved. Vaidya and Clifton [8] found that this works in data warehouses, but trades off privacy for accuracy. Another approach uses cryptographic protocols. Multiple sites can cooperate to learn global data mining results without revealing the individual records that each site has. Shaw and Li [9] presented an attribute analysis framework to de-identify personal health information for data mining. El Emam et al [10] has analyzed algorithms for "re-identifying" individuals based on combining "de-identified" data from distributed sources. Our concept of "Client Master Index" is taken from work on the Australian electronic health record initiative [11].

Privacy in identity management systems including the concept of a Circle of Trust (CoT) is discussed by Shin et al [12]. In a CoT, individual identity is protected by an Identity Provider, while allowing services within the CoT to share the individual's information in a manner that ensures the individual's permission is obtained and their identity protected. Koch and Möslein [13] explores identity management and privacy including a discussion of anonymous versus pseudonymous identity.

Peyton and Nozin [14] proposed an Information Transfer Registry (ITR) to support the auditing of information transfers between businesses in B2B. An analysis of privacy compliance within the Liberty Alliance framework is in Alsaleh and Adams [15]. The Liberty Alliance architecture is specified in [16, 17], while security and privacy are discussed in [18, 19] and the support for a unified data model is described in Kellomäki and Kainulainen [20]. Our ePrescription scenario is adapted from [5].

3 Liberty Alliance ePrescription scenario

Figure 1 shows a Liberty Alliance Circle of Trust. There is a prescription service, ePrescription that is used by doctors who write prescriptions for patients. Prescriptions are sent to the patient's pharmacy, ePharmacy, for fulfilment and the pharmacy bills the patient's insurance company, eInsurance. Throughout the scenario, the Identity Provider provides a single sign on (SSO) service so that users need to "log in" only once. After that, each service (ePrescription, ePharmacy, eInsurance) recognizes the patient by a different pseudonym known only to them which is provided by the Identity Provider through an identity mapping service. When a service wishes to access data about a user from another service, it first discovers that service, using a discovery service within the Identity Provider to obtain an end point reference (EPR). The EPR contains security tokens that allow the invoked service to extract their pseudonym for the user without revealing it to the calling service. The user must have granted permission for the two services to share the data. If not, the Identity Provider invokes an interaction service to contact the user and obtain their permission.

Here is a detailed description of the steps involved in the scenario:

1. A Doctor is redirected to the Identity Provider to sign on to the CoT the first time they attempt to access any service in the CoT.

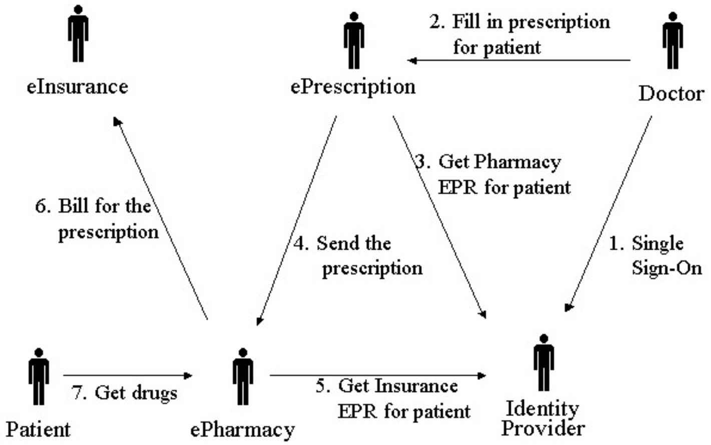

eInsurance
ePrescription
2. Fill in prescription for patient
Doctor
6. Bill for the prescription
3. Get Pharmacy EPR for patient
4. Send the prescription
1. Single Sign-On
Patient
7. Get drugs
ePharmacy
5. Get Insurance EPR for patient
Identity Provider

Figure 1: Liberty Alliance ePrescription scenario.

2. The Doctor accesses the ePrescription service, selects a patient and enters a prescription. The Patient has previously given permission for the Doctor to access their information. The Doctor is recognized by ePrescription using the pseudonym passed to it by the Identity Provider.
3. The ePrescrption Service communicates with a discovery service in the Identity Provider to obtain an end point reference (EPR) that enables it to communicate with ePharmacy on behalf of the Patient.
4. The ePrescription service sends the prescription to the ePharmacy using the EPR. The EPR contains security tokens from which ePharmacy can extract its pseudonym for the Patient.
5. The ePharmacy extracts its pseudonym for the Patient from the EPR, and uses the discovery service of the Identity Provider to obtain another EPR that will enable it to invoke the Patient's eInsurance service
6. The ePharmacy sends a claim to eInsurance using the second EPR. The second EPR contains security tokens from which eInsurance extracts its pseudonym for the Patient.
7. The Patient identifies themselves to the ePharmacy (by authenticating with the Identity Provider) and receives their prescription drugs.

4 Knowledge discovery process in a circle of trust (CoT)

One would like to monitor an on-line prescription drug service for signs of misuse through a knowledge discovery process. This "misuse" could be criminally fraudulent activity in which prescriptions for prescription narcotics are obtained fraudulently in order to resell the drugs on the black market. Or the "misuse" could simply indicate a need for education, for example if there are new guidelines to control the prescription of antibiotics in order to reduce the likelihood of a "super bug" outbreak that is resistant to antibiotics

In this section, we use our ePrescription scenario as an example to examine the special requirements that must be addressed within a Circle of Trust (CoT) at each stage of a knowledge discovery process based on the CRISP-DM model [6]. In particular, we have found it necessary to add a seventh stage, "data collection", in the CRISP-DM model, to address the issues of assembling data from different organizations in a B2B network.

4.1 Business understanding

The first stage focuses on understanding the project requirements from a business perspective, and converting this knowledge into a defined data mining problem and project plan. The two main requirements specific to a CoT that must be addressed from a business point of view is to understand the business relationships that must be established to mine data from different organizations within the CoT and to ensure the privacy rights of individuals are protected.

In our scenario, data from each of ePrescription, ePharmacy and eInsurance may be relevant to detecting signs of misuse, such as a doctor prescribes more than other doctors; a patient receives more drugs than other patients; a pharmacy dispenses more drugs than other pharmacies; a health insurance company pays for more prescription than other health insurance. As well, there are attributes that cut across services between what was prescribed, fulfilled and paid for, along with demographic information that could be correlated to potential misuse.

To allow data mining process to take place, there must be business agreements in place for the three services to provide their audit data and permission to use the data for such purposes must be explicitly or implicitly be given according to the relevant privacy legislation. Explicit consent requires informing the user that their data is being shared and obtaining their signed permission. If they refuse, their data cannot be included. Implicit consent requires that the knowledge discovery process is arguably part of the service provided and therefore the user's consent is implied by the fact they requested the service. If the data is de-identified or shared at an aggregate level, it can also be shared. In summary:

1. Establish business agreements with services to obtain relevant data
2. Ensure sharing of data complies with relevant legislation through explicit consent, implicit consent, or de-identification of data.

4.2 Data understanding

The data understanding stage proceeds with activities to answers what is the data, where to find it, what is the format, and identify data quality problems. The two main requirements specific to a CoT that must be addressed in this stage is the mechanism to integrate and link data from different organizations and the technique to handle identifiable data.

In our scenario each service will have a different pseudonym to refer to the same user, and each service will have its own set of attributes that it stores related to events which take place in the operations of the CoT. The required attributes from each of ePrescription, ePharmacy and eInsurance will have to be identified, obtained and linked in some manner before further analysis is possible. The infrastructure for collecting attributes from different attribute providers and linking pseudonyms without compromising identity will be discussed in more detail in section 4.3 and section 5.

In addition, the attributes and data collected will have to reviewed to ensure that the combination of attributes needed for data mining are not potentially identifiable, for example, if geographic attributes like postal code are combined with other attributes. Recent work in El Emam et al [10] gives a comprehensive treatment of the issues. In summary:

- Identity attributes required from different services and linkages.
- Ensure attributes in combination do not potentially violate privacy.

4.3 Data collection

The data collection stage is a new stage, we introduce to address how to integrate attributes from distributed services and link them via pseudonym mapping without compromising identity. Figure 2 diagrams the main components required

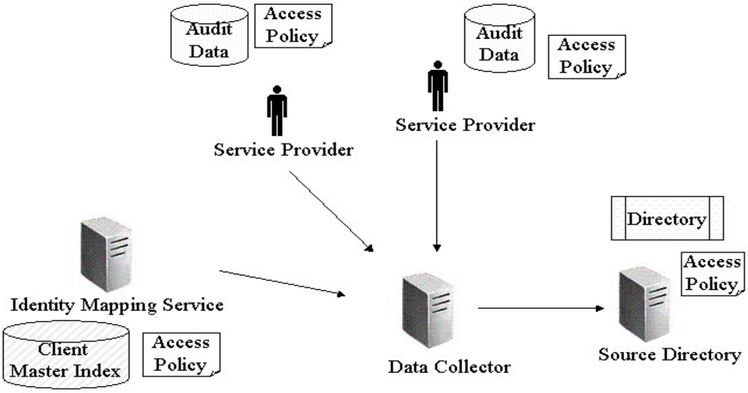


Figure 2: Trusted data collection architecture.

for data collection in a CoT. The Data Collector, and Source Directory are new services we propose; otherwise all other components are part of the Liberty Alliance framework for a CoT.

Each Service Provider has a database of attributes with respect to the services it provides as well as access policies that control who can access those attributes. The Source Directory manages a directory of attributes available from each Service Provider based on a common data model across the CoT as defined in Kellomäki [20] that the Data Collector can use to formulate queries for result sets from each Service Provider. The Data Collector acts on behalf of a data mining user or service that must have the required access to obtain the result sets. The result sets returned are indexed by pseudonyms particular to each Service Provider, but by interacting with the Identity Management Service, the Data Collector maps those pseudonyms to a single set of pseudonyms specific to it. Attributes for a given user from different Service Providers are linked without compromising identity (discussed in detail in section 5). The Identity Mapping Service of the Identity Provider contains the master list of pseudonyms (but no other data) which can serve as a Client Master Index, as defined in Fitzgerald [11] for linking distributed attributes. In summary, for data collection:

1. Define the results sets and attributes from different service providers for which the Data Collector will retrieve, link and resolve pseudonyms.
2. Ensure that the data mining user or service has the required access permissions at each service provider to obtain the required data.

4.4 Data preparation

The goal of data preparation stage is to produce datasets suitable for mining from the raw data collected in the data collection stage. Tasks include select data, clean data, construct data, and integrate data and transform data. This stage proceeds as it would in a generic knowledge discovery stage once the issues around distributed sources for attributes and resolution of pseudonyms specific to a CoT are resolved in the data collection stage.

4.5 Modelling

In modelling stage, various techniques are selected and applied and their parameters are calibrated to optimal values depending on the specific knowledge discovery task. There are many approaches that could apply to prescription misuse. The main issue that may need to be addressed specific to a Circle of Trust is the issue of privacy. Although, explicit links to identity are protected by pseudonyms in a Circle of Trust, techniques such as randomization and perturbation in Vaidya and Clifton [8] may needed to ensure that the combination of attributes collected do not result in identifiable data sets. The need for such techniques will have been identified in the Data Understanding stage. In summary,

1. If issues with identifiable data sets were identified in the Data Understanding stage, ensure that appropriate techniques and algorithms are used to resolve them.

4.6 Evaluation

In the evaluation stage, it is important to more thoroughly evaluate the model and review the steps executed to mine the data to be certain it properly achieves the business objectives before proceeding to deployment. In monitoring misuse, have all appropriate business agreements and privacy legislation been adhered to? Will the results have business, ethical or legal implications and who will receive this information and who will be responsible for acting on the information in what manner? In summary, the following should be verified:

1. Appropriate consent is obtained from users for the execution of the model on their data.
2. Business agreements are in place both for obtaining the data from service providers and sharing the results of executing the model.
3. The data mining users or services have the required access rights at each service provider for the data required.
4. The results of executing the data model will not be identifiable (compromising privacy).
5. The inferences made by the model will be statistically sound, and that all potential business and legal implications of the results the model may generate have been evaluated.

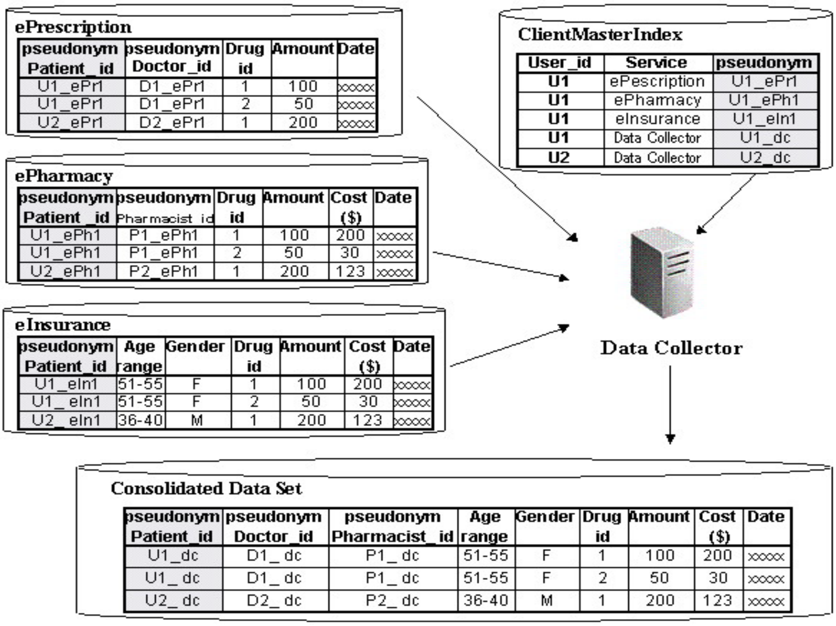
4.7 Deployment

In this stage, the knowledge discovery process is executed, and the results are packaged and communicated. It includes both the knowledge extracted from the data and how it is communicated as well as feedback from the process and experience of mining the data. There are no specific extra requirements for a Circle of Trust, other that dealing with the issues around privacy and the cooperation of participating organizations that have already been mentioned. It is possible that the results of the data mining can be made available as a service within the CoT, in which case the usual constraints and architecture of the Liberty Alliance framework for any service provider would apply.

5 Trusted data collection for ePrescription scenario

The trusted data collection architecture is the essential element for establishing a knowledge discovery process in a Circle of Trust (CoT). It enables one to define the results sets and attributes from different service providers for which the Data Collector will retrieve, link and resolve pseudonyms. Figure 3, illustrates how the attributes for the ePrescription scenario are distributed throughout the CoT. The doctor_id is only available from ePrescription. The pharmacist_id is only available from ePharmacy, and the demographic information about the patient (age, gender) is only available from eInsurance. There is also duplication of attributes in the data each service has: patient_id, drug_id, amount, cost, date. Amount and cost can be used as an accounting check against fraud, whereas patient_id, drug_id and date can be used to link the data from different sources into a single dataset. However, patient_id is a different pseudonym at each service which is meaningless outside the service (same with doctor_id and pharmacist_id).

The Identity Mapping Service maintains a Client Master Index to resolve pseudonyms. When the Data Collector collects the different result sets from each service, it will interact with the Identity Mapping Service to convert each pseudonym into a single Data Collection pseudonym. In doing so, it is able to resolve the patient_id pseudonyms to link the data for user U1 into a single record indexed by U1_dc. Similarly Doctor1 and Pharmacist1 are resolved to Doctor1_dc, Pharmacist1_dc. In this manner, the identity of U1, Doctor1, and Pharmacist1 are protected, but a consolidated view across all services is created.



ePrescription

pseudonym Patient _id	pseudonym Doctor_id	Drug id	Amount	Date
U1_ePr1	D1_ePr1	1	100	xxxxx
U1_ePr1	D1_ePr1	2	50	xxxxx
U2_ePr1	D2_ePr1	1	200	xxxxx

ClientMasterIndex

User_id	Service	pseudonym
U1	ePescription	U1_ePr1
U1	ePharmacy	U1_ePh1
U1	eInsurance	U1_eIn1
U1	Data Collector	U1_dc
U2	Data Collector	U2_dc

ePharmacy

pseudonym Patient _id	pseudonym Pharmacist_id	Drug id	Amount	Cost ($)	Date
U1_ePh1	P1_ePh1	1	100	200	xxxxx
U1_ePh1	P1_ePh1	2	50	30	xxxxx
U2_ePh1	P2_ePh1	1	200	123	xxxxx

eInsurance

pseudonym Patient_id	Age range	Gender	Drug id	Amount	Cost ($)	Date
U1_eIn1	51-55	F	1	100	200	xxxxx
U1_ eIn1	51-55	F	2	50	30	xxxxx
U2_ eIn1	36-40	M	1	200	123	xxxxx

Consolidated Data Set

pseudonym Patient_id	pseudonym Doctor_id	pseudonym Pharmacist_id	Age range	Gender	Drug id	Amount	Cost ($)	Date
U1_dc	D1_ dc	P1_ dc	51-55	F	1	100	200	xxxxx
U1_ dc	D1_ dc	P1_ dc	51-55	F	2	50	30	xxxxx
U2_ dc	D2_ dc	P2_ dc	36-40	M	1	200	123	xxxxx

Figure 3: Data collection for ePrescription scenario.

Should eMonitor discover patterns of behaviour that require action there are a number of options. If for example, it seemed that Doctor1 was not following the new accepted protocol for prescribing antibiotics; educational material could be communicated via an interaction service that would notify Doctor1 within the CoT based on an interaction pseudonym. If there was a persistent pattern of behaviour not in keeping with the policies of the CoT, Doctor1 could be rejected access by the Identity Provider. In the case of criminal activity, the matter could be referred to the police, who might be able to obtain a court order which would require the Identity Provider to provide access to the individual.

6 Conclusions

This paper proposes an additional stage, data collection, for privacy preserving distributed data mining based on CRISP-DM standard [6] and outlines a trusted data collection architecture to support data collection within the Liberty Alliance federated identity management framework. The approach ensures that identity can be protected by a system of pseudonyms and still allow data about individuals to be correlated from different organizations within a Circle of Trust into a single consolidated data set. While the pseudonyms protect direct identification, there is still the well known problem that the correlated data in aggregate may be potentially identifiable. Known approaches for addressing this issue would have to be applied when working with the consolidated data set.

References

[1] European Union Directive on Privacy and Electronic Communications. European Parliament, Brussels, Belgium, 2002. http://register.consilium.eu.int/pdf/en/02/st03/03636en2.pdf

[2] HIPAA, Health Insurance Portability and Accountability Act, United States Congress, United States, 1996. http://aspe.hhs.gov/admnsimp/pl104191.htm

[3] PIPEDA, The Personal Information Protection and Electronic Documents Act, Department of Justice, Canada, 2000. http://laws.justice.gc.ca/en/P-8.6/text.html

[4] The Liberty Alliance, www.projectliberty.org/

[5] Benthin, T., Liberty ePrescription Scenario ver. 1.1, Liberty Alliance Project, 2005. www.projectliberty.org/liberty/adoption/healthcare

[6] Shearer, C., The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of Data Warehousing*, 5(4), pp. 13-22, 2000.

[7] Zaidi, S., Abidi, S. & Manickam, S., Distributed Data Mining From Heterogeneous Healthcare Data Repositories: Towards an Intelligent Agent-Based Framework, *Proc. of the 15 the IEEE Symposium on Computer-Based Medical Systems,* 2002.

[8] Vaidya, J. & Clifton, C., Privacy-Preserving Data Mining: Why, How, and What For? *IEEE Security & Privacy*, New York, NY, Nov./Dec., 2004.

[9] Shaw, M. & Li, J., Protection of Health Information in Data Mining. *Int. Journal of Healthcare Technology and Management*, 6(2), 2004.

[10] El Emam, K., Jabbouri, S., Sams, S., Drouet, Y. & Power, M., Evaluating Common De-Identification Heuristics for Personal Health Information, *Journal of Medical Internet Research*, 8(4):e28, 2006.

[11] Fitzgerald, P., (eds). *HealthConnect* Interim Research Report, Volume 2, Australia, 2003. www.health.gov.au/internet/hconnect/publishing.nsf/ Content/43598FE37A3E7270CA257128007B7EB7/$File/v2.pdf

[12] Shin, D., Ahn, G-J & Shenoy. P., Ensuring Information Assurance in Federated Identity Management, *IEEE Intl. Conference on Performance, Computing, and Communications*, pp. 821-826, 2004.

[13] Koch, M. & Möslein, K.M., Identity Management for Ecommerce and Collaborative Applications. *International Journal of Electronic Commerce*, 9(3), pp. 11–29, 2005.

[14] Peyton, L. & Nozin, M., Tracking Privacy Compliance in B2B Networks, *Proc. Of Sixth International Conference on Electronic Commerce*, Delft, The Netherlands, 2004.

[15] Alsaleh, M. & Adams, C., Enhancing Consumer Privacy in the Liberty Alliance Identity Federation and Web Services Frameworks. *Proc. of the 6th Workshop on Privacy Enhancing Technologies (PET 2006)*, Cambridge, United Kingdom, June 2006

[16] Kemp, Y., (eds). Liberty ID-WSF Web Services Framework Overview, Liberty Alliance Project, 2004, www.projectliberty.org/liberty/resource_ center/papers

[17] Wason, T., (eds). Liberty ID-FF Architecture Overview, version 1.2, Liberty Alliance Project, March 2003. www.projectliberty.org/ liberty/resource_center/papers

[18] Hodges, J., Aarts R., Madsen P & Cantor, S., (eds). Liberty ID-WSF Authentication, Single Sign-On, and Identity Mapping Services Specification Ver2.0, Liberty Alliance Project, New Jersey, 2006. www.projectliberty.org/liberty/content/download/871/6189/file/liberty-id wsf-authn-svc-v2.0.pdf

[19] Landau, S., (eds). Liberty ID-WSF Security & Privacy Overview, version 1.0, Liberty Alliance Project, 2003, www.projectliberty.org/ resource_center/specifications/liberty_alliance_id_wsf_2_0_specifications

[20] Kellomäki, S. & Kainulainen, J., (eds). Liberty ID-WSF Data Services Template ver.2.1, Liberty Alliance Project, New Jersey, 2006. www.projectliberty.org/liberty/content/download/879/6213/file/liberty-idwsf-dst-v2.1.pdf