Shortcomings of Current Grid Middlewares Regarding Privacy in HealthGrids

Yassene MOHAMMED[1, *], Ulrich SAX [*], Fred VIEZENS, Otto RIENHOFF
Department of Medical Informatics, University of Goettingen, Germany

Abstract: Although grid computing middlewares are in research use since many years, they lack of particular security features for biomedical applications. The analysis of the common Globus middleware reveals several security-related shortcomings. As a result, extended security measures for HealthGrids have been identified. They include tools for auditing, tracking, fine grained access control for structured documents, trust and trust delegation. The German MediGRID project is facing this with an "Enhanced Security" package intending to bridge the gap between current legal, data protection as well as data security requirements and the available grid technology.

Keywords: Data Protection, Privacy, Security, GSI, GT4, Enhanced Security

Introduction

There are several challenges the biomedical community has to face until biomedical grids will be largely in use. Beyond the problem of retrieving the relevant data sets using the metadata description, data access control is of paramount importance, as the owners of the data are foremost patients. Due to the heterogeneity of the data an additional ontology process is needed to homogenize the data. Figure 1 shows the grid data-flow for biomedical applications differing from usual grids by the need of retrieval, authorization and homogenizing steps.

In contrast to conventional grid applications, medical applications typically use high dimensional data. Biomedical data are not only heterogeneous; rather they contain different information types and different levels of privacy. They vary from aggregated data describing population and diseases (epidemiology, clinical practice, clinical trials), to more granular patient data and pathological descriptions (health record, clinical history, physical exams) and to cellular and molecular data (histology, genetic test results and genomic data) [1-3]. Given semantic data interoperability, the researcher can correlate and analyze the data using suitable biomedical informatics methods and tools. On the other hand having this data online with the suitable tools to correlate, merge and analyze creates new challenges for data protection and data security [4].

[1] Corresponding Author: Yassene Mohammed, Department of Medical Informatics, Georg-August-University of Goettingen, 37075 Goettingen, Germany. E-Mail: ymohammed@med.uni-goettingen.de.

[*] Authors contributed equally.

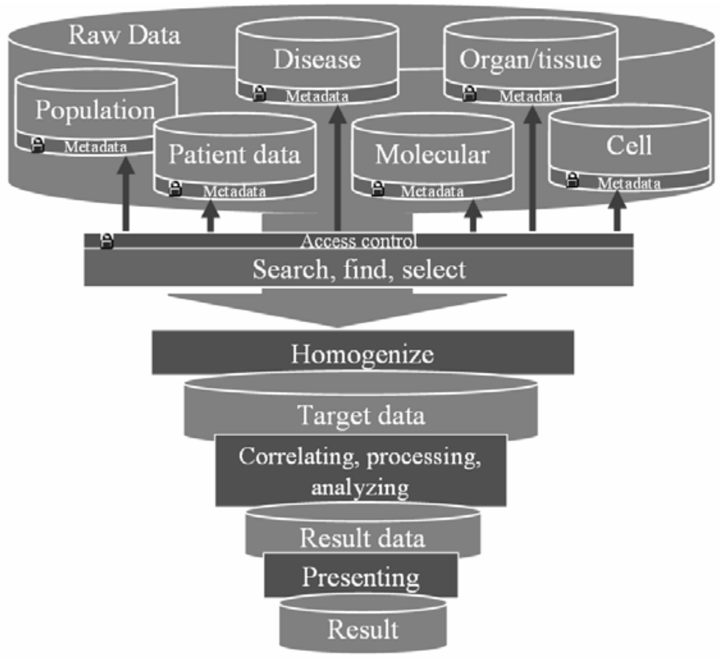

Figure 1. Data flow in MediGRID as an example for a HealthGrid: numerous data formats as well as high dimensional data in medical applications are the rational for an additional homogenizing step, before the usual eScience data processing can be started.

1. Methods

In order to analyze the privacy needs we examined the current security in grid middleware focusing on the grid security infrastructure in the Globus Toolkit. Although there are quite some grid middlewares like gLite [5] and UNICORE [6], the Globus Toolkit 4 (GT4) is widely used and could be considered as the "standard" Grid middleware for biomedical research. The security tools in GT4 deal with [7]:

- authentication: establishing the identity of users or services,
- communication security
- authorization: determining who is allowed to perform what actions, and
- other supporting functions such as managing user credentials and maintaining group membership information.

GT4 provides distinct web services (WS) and pre-WS authentication and authorization capabilities [7-10]. Both use standard X.509 certificates and proxy certificates [11], which are used to identify persistent entities such as users and servers and to support the temporary delegation of privileges to other entities.

Following the Globus design model, which intends to use current internet technologies with as less modifications as possible and the hour glass model for new standards [12], the Globus Security Team implements security as a "five layers grid security infrastructure (GSI)" [8] (see Table 1) based on standard X.509 certificates.

Table 1. The five layers in Grid Security Infrastructure (GSI) as presented in GT4 [8]

Authorization	Grid-Mapfile/ SAML(Security Assertion Markup Language)
Delegation	X.509 Proxy Certificates
Authentication	X.509 ID Certificates
Message	WS-Security/ WS-SecureConversation
Message Format	SOAP (Simple Object Access Protocol)

The Grid Security Infrastructure (GSI) builds the core for security in the GT4 middleware. The use of this security infrastructure in combination with job submission, data management, and execution management contrives the secure Grid infrastructure. In this context, GT4 provides Data Management tools [13] for

- Data Movement including GridFTP and Reliable File Transfer (RDT),
- Data Replication including Replica Location Services (RLS), and
- Higher Level Data Services -Data Replication Services (DRS).

These tools are designed to work in combination with the GSI, which leads indeed to suitable confidentiality of communication and to data integrity required for networks for biomedical research - HealthGrids. In Figure 2, this fulfills the data security requirements for the first step – the Upload service.

In contrast to "uploading" data in the Grid, the second step – retrieval – requires more comprehensive and advanced data management. To some degree this could be introduced by available "plugins" designed to work with GT4. With tools like Storage Resource Broker (SRB) – a data grid management system – [14, 15] and Data Access and Integration Services (OGSA-DAI) [16-18] one can achieve the necessity of data availability.

While web services provide the ability to access and manipulate data, there is a need to define conventions for managing data. This led to the development of the WS-Resource Framework (WSRF) [19, 20]. WSRF, Grid Resource Allocation Management (GRAM) [21] and Monitoring and Discovery System (MDS) [22], representing the execution and information management in GT4, provide confidentiality within applications.

Some HealthGrid projects, namely the French MEDIGRID [23], implemented their own light weighted "µgrid" middleware [24, 25], suitable security for this middleware - "Sygn" [26] and an encrypted storage mechanisms of medical data on grids [27]. The aspects of fine grained authorization with respect to user-organization relationship were discussed and implemented in Sygn. Sygn was designed to be more efficient than the Community Authorization Service (CAS) developed by the Globus team [28] and than the Virtual Organization Membership Service (VOMS) [29]. MammoGrid project [30] handled security as a service 'on the Grid' and build it on the top of the GT4-GSI tools [31, 32]. GEMSS project [33] considers security for the case of medical simulation and image processing on the grid and reflects in the implementation the legal security requirements [34, 35].

While most grid projects follow the common grid middleware in focusing on security, less work has been done regarding data protection in grids. A legal framework for the protection, security and transport of personal data as well as patient data is introduced in different EU directives. E.g. directive 95/46/EC concerns processing of personal data and free movement of such data, directive 97/66/EC regards the protection of privacy in the telecommunications sector, directive 99/93/EC describes a framework for electronic signatures, and directive 2002/58/EC deals with privacy and electronic communications. Country specific implementations vary among the EU countries [36].

The legal framework implies special requirements regarding data security and data protection [36-38] already beeing included in most grid middlewares: (1) Confidentiality of communication and application, (2) Integrity and authenticity, (3) Data availability, and (4) Personal responsibility of data processing.

Additional data protection requirements arise dealing with personal data especially in the health care sector:

- Data necessarity principle: disclose all medical and medical-relevant data of a patient, but not more than needed data for the treatment and provision of that patient.
- Context of treatment: medical and medical-relevant data of a patient should be disclosed only to the personals participated in his treatment and only the information related to this treatment is allowed to be disclosed.
- Patient consent: the patient should formally agree on the storage of his medical and medical-relevant data
- The guarantee of patient rights: the possibility of rectification, blocking, deletion of his personal data should be presented.

Offering services to fulfill these requirements on HealthGrids helps the developer to implement her application according to the legal data protection and security level.

2. Results

As a Result of this Analysis, extended security measures for HealthGrids have been identified. Beyond anonymization and pseudonymization, which are procedures to be accomplished before uploading sensitive/patient data (see Figure 2 first step), the above mentioned technologies fulfill to a good degree the requirement for data security in the grid. Anyhow we still need to know who did what when and why, namely to follow the responsibilities on the grid in order to completely fulfill the legal data security and protection requirements. On the other hand, and especially because grid middlewares are yet developed not for the special use by the biomedicine community, the requirements of data protection should be considered as well.

Several security extensions have been discussed in MediGRID [39], the biomedicine community grid project in the German national grid infrastructure D-Grid [40] funded by the Federal Ministry of Education and Research (BMBF). Our analysis shows the essential “Enhanced Security” elements for a HealthGrid:

- **Auditing** (a posteriori): an audit trail consists of log files and activity protocols. Auditing is crucial for any privacy regulation assessment. Beyond the relevant user and machine data, especially valid time stamps and time periods are needed for an efficient audit. Further dimensions of auditing are data provenance and data annotation.

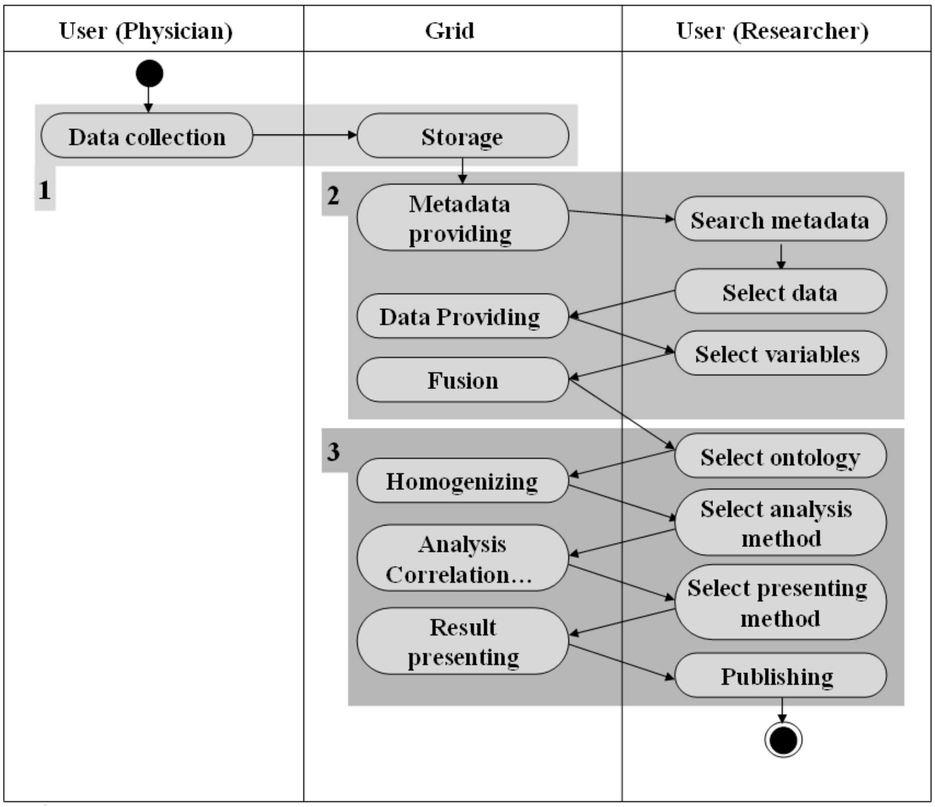

Figure 2. Activity diagram of the service flow in MediGRID as an example for a HealthGrid: 1- Upload on the grid 2- Retrieval: the user (researcher) can retrieve and select the data he needs for his work or research - the researcher prepares the data for processing, anyhow the data it self is not changed yet, 3 – Processing: here the researcher will use algorithms and processing power available on the grid to process and analyze the data intending to receive the needed results.

- **Trackability** (a priori): Additionally to auditing, trackability requires knowledge about where transfers, transactions, calculations and storage of person related data take place. This has to be part of the informed consent process between doctor and patient, as the patient data leave the doctor-patient confidentiality.
  → Auditing and tracking-possibilities cover the requirement to retrace responsibilities and retain the separation of identification data and medical data in order to preserve anonymity or pseudonomity.
- **Access rights and control**: in addition to the authentication and authorization, biomedical grid computing needs fine grained access control with respect to access rights within medical documentations, which means, that the current access control on file level (e.g. Grid-Mapfiles) not suffice, as structured medical documents [41, 42] provide different sections with a different degree of confidentiality.
- **Confidentiality**: In addition to fine grained access control in structured documents, fine grained confidentiality services have to be modified for grid computing.

→ Fine grained access rights and control as well as fine grained confidentiality fulfill the requirements of releasing only necessary data and retain the doctor-patient confidentiality.

- **Trust and trust-delegation:** trust relations and delegation as well as trust hierarchies from every day life have to be set up electronically. Using the data of a minor or a person with dementia requires that an authorized person signs electronically on behalf of those persons (eConsent). These workflows are described in some projects [43, 44], but have to be adapted for grid usage.
- **Safety:** security of data in possibly dynamic grid environments requires policies for data storage and policies for data management.
  → Safty reflects the need to develop and adopt suitable policies for the use and storage of data; a complementary safeguard principle when intending to use sensitive data considering the availability concept in time (long term archiving) and place (replicas).

The elements of the Enhanced Security consider after all the current requirements of data protection and data security intending to make grid technologies better suitable for the biomedicine community. In the future, Enhanced Security should be also flexible to fulfill future legal requirements and new developments in the medical area, e.g. genome wide association studies.

3. Conclusion and Outlook

The development of standards for data protection and data security in grids is crucial for the success of grid computing in many grid communities. Current grid middlewares lack standards and have technological shortcomings in regard to fulfill basic data protection and data security requirements. The need for a secure grid is not only an issue of computing in biomedicine. Within the German D-Grid communities there is a notable interest in the different security aspects especially in the automotive sector concerning intellectual property protection. Meanwhile the “classic” grid communities - for example climate researchers - aim for similar security standards as well. This means a long development process until biomedical and intellectual property related grid computing can make full use of the grid [4, 45, 46].

The Enhanced Security package in MediGRID is rather a one step towards enabling grid technology to be used by the biomedicine community than a complete solution. In biomedicine applications sustainability should be guaranteed. That means we need to deal with two further dimensions for a more suitable solution: future development of the grid technologies and legal framework, and international collaborative work on the country specific (legal) requirements.

The 26th international conference on privacy and data protection in Wrocław 2004 resulted in a resolution about a „Privacy Framework Standard“. The resolution urges the International Standards Organzation (ISO) to work on privacy and Data Protection standards: „Development from Privacy Law into Privacy Standards”. The “Privacy Enhancing Technologies” (PET) [47, 48] are of interest for the future ISO Privacy-Standard [49]. This has to be closely monitored in the interest of the biomedical grid community in order to set up a sustainable grid infrastructure.

Each change in the legal framework or in the technology in regard to grid-computing use by the biomedical community should take these standards into account. A “converging” between the legal framework and the technical solutions of data protection and data security to the common ISO privacy standards should be considered [4]. As it is not expected to have them before 2008 [49], we need to keep track of the development of the ISO privacy standards in order to keep the converging time later as short as possible.

Acknowledgment

This work was supported by the e-Science research program of the Federal Ministry of Education and Research (BMBF) under the promotion number: 01AK803H

References

[1] Martin-Sanchez, F., V. Maojo, and G. Lopez-Campos, *Integrating genomics into health information systems*. Methods Inf Med, 2002. **41**(1): p. 25-30.

[2] Sax, U. and Y. Mohammed. *Data-related Challenges of Genotype - Phenotype studies*. in *GMDS 2006 - Genome Wide Association Studies in Complex Traits*. 2006. Leipzig.

[3] Sax, U. and S. Schmidt, *Integration of genomic data in Electronic Health Records--opportunities and dilemmas*. Methods Inf Med, 2005. **44**(4): p. 546-50.

[4] Mohammed, Y., Viezens, F., Sax, U., Rienhoff, O. , *Rechtliche Aspekte bei Grid-Computing in der Medizin*, in *Rechtliche Aspekte der Telemedizin*, W. Niederlag, C. Dierks, and H.U. Lemke, Editors. 2006. p. 235-245.

[5] *gLite*. [2006 Dec. 27]; Available from: http://glite.web.cern.ch/glite/.

[6] *UNICORE*. [2006 Dec. 27]; Available from: www.unicore.org.

[7] The_Globus_Security_Team. *GT 4.0 Security*. 2006 [2006 Dec. 27]; Available from: www.globus.org/toolkit/docs/4.0/security/.

[8] Foster, I. and C. Kesselman, *The grid : blueprint for a new computing infrastructure*. 2nd ed. The Elsevier series in grid computing. 2004, Amsterdam ; Boston: Morgan Kaufmann. xxvii, 748 p.

[9] The_Globus_Security_Team. *Globus Toolkit Version 4 Grid Security Infrastructure: A Standards Perspective*. 2005 [2006 May 31]; Available from: http://www.globus.org/toolkit/docs/4.0/security/GT4-GSI-Overview.pdf.

[10] Welch, V., et al. *Security for Grid Services*. in *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC'03)*. 2003 IEEE Computer Society

[11] IETF. *Internet X.509 Public Key Infrastructure (PKI), Proxy Certificate Profile*. 2004; Available from: www.ietf.org/rfc/rfc3820.txt.

[12] The_Globus_Security_Team. *GT 4.0 Security: Key Concepts*. 2006 [2006 Dec. 27]; Available from: www.globus.org/toolkit/docs/4.0/security/key-index.html

[13] The_Globus_Alliance. *GT 4.0 Data Management* [2006 Dec. 27]; Available from: www.globus.org/toolkit/docs/4.0/data/.

[14] Rajasekar, A., et al., *Storage Resource Broker - Managing Distributed Data in a Grid.* Computer Society of India Journal, 2003. **33**(4): p. 42-54

[15] San_Diego_Supercomputer_Center_(SDSC). *Storage Resource Broker* 2006 [2006 Dec. 27]; Available from: www.npaci.edu/DICE/SRB/.

[16] Karasavvas, K., et al., *Introduction to OGSA-DAI Services*, in *Scientific Applications of Grid Computing, First International Workshop*. 2005, Springer Berlin / Heidelberg: Beijing. p. 1-12.

[17] Herrero, P., M.S. Pérez, and V. Robles, *Scientific applications of grid computing : first international workshop, SAG 2004, Beijing, China, September, 20-24, 2004 : revised selected and invited papers*. Lecture notes in computer science, 3458. 2005, Berlin ; New York: Springer. x, 208 p.

[18] Open_Middleware_Infrastructure_Institute_UK_(OMII-UK). *Open Grid Services Architecture - Data Access and Integration Services*. [2006 Dec. 27]; Available from: www.ogsadai.org.uk/.

[19] The_Globus_Alliance, IBM, and HP. *WS-Resource Framework (WSRF)*. 2004 [2006 Dec. 27]; Available from: www.globus.org/wsrf/.

[20] OASIS_Web_Services_Resource_Framework_(WSRF)_Technical_Committee. *Web Services Resource Framework (WSRF)*. 2006 [2006 Dec. 27]; Available from: www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf.

[21] The_Globus_Alliance. *GT Execution Management: Grid Resource Allocation Management (GRAM)*. [2006 Dec. 27]; Available from: www.globus.org/toolkit/gram/.

[22] The_Globus_Alliance. *GT Information Services: Monitoring & Discovery System (MDS)*. [2006 Dec. 27]; Available from: www.globus.org/toolkit/mds/.

[23] *MEDIGRID*. [2006 Dec. 27]; Available from: www.creatis.insa-lyon.fr/MEDIGRID/.

[24] Montagnat, J., V. Breton, and E.M. I, *Partitioning medical image databases for content-based queries on a Grid.* Methods Inf Med, 2005. **44**(2): p. 154-60.

[25] MEDIGRID. *µgrid*. 2005 [2006 dec. 28]; Available from: www.creatis.insa-lyon.fr/MEDIGRID/.

[26] Seitz, L., et al., *Authentication and Authorisation Prototype on the µgrid for Medical Data Management*, in *From Grid to Healthgrid - Proceedings of Healthgrid 2005*, T. Solomonides, et al., Editors. 2005, IOS Press: Amsterdam ; Washington, DC. p. 222 - 233.

[27] Seitz, L., J.M. Pierson, and L. Brunie, *Encrypted storage of medical data on a grid.* Methods Inf Med, 2005. **44**(2): p. 198-201.

[28] Pearlman, L.W., V. Foster, I. Kesselman, C. Tuecke, S. , *A Community Authorization Service for Group Collaboration*, in *Third International Workshop on Policies for Distributed Systems and Networks, June 5-7, 2002, Monterey, California, USA : proceedings*, IEEE Computer Society. TC on Distributed Processing., Naval Postgraduate School (U.S.), and United States. Office of Naval Research., Editors. 2002, IEEE Computer Society: Los Alamitos, Calif. p. 50-59.

[29] R. Alfieri, R.C., V. Ciaschini, L. dellAgnello, Á. Frohner, A. Gianoli, K. Lõrentey, F. Spataro3, *VOMS, an Authorization System for Virtual Organizations*, in *Grid Computing - First European Across Grids Conference*, G. Goos, J. Hartmanis, and J.v. Leeuwen, Editors. 2004, Springer Berlin / Heidelberg. p. 33-40

[30] MammoGrid, *www.mammogrid.com*.

[31] Estrella, F., et al., *Experiences of engineering Grid-based medical software.* Int J Med Inform, 2006.

[32] Estrella, F., R. McClatchey, and D. Rogulin, *The MammoGrid Virtual Organisation - Federating Distributed Mammograms.* Stud Health Technol Inform, 2005. **116**: p. 935-40.

[33] GEMSS, *www.gemss.de*.

[34] Herveg, J.A.M., et al. *GEMSS: Privacy and security for a Medical Grid*. in *HealthGRID*. 2004. Clermont-Ferrand, France.

[35] Herveg, J.A.M. and Y. Poullet. *Directive 95/46 and the use of GRID technologies in the heathcare sector: selected legal issues*. in *Healthgrid 2003*. 2003. Lyon.

[36] Sax, U., Mohammed, Y., Viezens, F., Rienhoff, O., *Grid-Computing in der Biomedizinischen Forschung – Datenschutz und Datensicherheit*. Medizinische Informatik, Biometrie und Epidemiologie. Vol. 90. 2006, München: Medizin und Wissen Publ.Comp.

[37] Sax, U., *Modellierung ausgewählter Sicherheitsaspekte der "ambulant-stationären Verzahnung" im deutschen Gesundheitswesen (Dissertation)*, in *Mathematisch-Naturwissenschaftliche Fakultäten*. 2002, Georg-August-Universität Göttingen: Göttingen. p. 126.

[38] van_der_Haak, M., *Architekturkonzepte für einrichtungsübergreifende elektronische Patientenakten am Beispiel des Tumorzentrums Heidelberg/Mannheim*, in *Medizinische Biometrie und Informatik*. 2006, Universitaetsklinkum Heidelberg: Heidelberg. p. 231.

[39] MediGRID, *www.medigrid.de*. 2005.

[40] D-Grid, *www.d-grid.de*. 2005.

[41] Dolin, R.H., et al., *HL7 Clinical Document Architecture, Release 2.* J Am Med Inform Assoc, 2006. **13**(1): p. 30-9.

[42] HL7, *HL7 Receives ANSI Approval of Three Version 3 Specifications Including CDA, Release 2*. 2005.

[43] TMF. *Telematikplattform für Medzinische Forschungsnetze*. 2006 [2006 Dec. 28.]; Available from: www.tmf-ev.de.

[44] TMF. *AG Datenschutz*. 2006 [2006 March 16.]; Available from: www.tmf-ev.de/site/DE/int/AG/DS/container_ag_ds.php.

[45] Rienhoff, O., *The SIREN legal workshops: list of urgent legal actions for telemedicine.* Stud Health Technol Inform, 1999. **64**: p. 61-4.

[46] Rienhoff, O., *A legal framework for security in European health care telematics*. Studies in health technology and informatics, v. 74. 2000, Amsterdam ; Washington, DC: IOS Press. x, 192 p.

[47] Hes, R. and J.J. Borking, *Privacy-enhancing technologies : the path to anonymity*. Rev. ed. Achtergrondstudies en verkenningen ; 11. 1998, The Hague: Registratiekamer. 54 p.

[48] Federrath, H. *Privacy Enhanced Technologies: Methods – Markets – Misuse* in *Trust, Privacy and Security in Digital Business* 2005. Copenhagen, Denmark: Springer Berlin / Heidelberg.

[49] Borking, J.J. *Privacy Standards for Trust*. in *27th International Conference on Privacy and Personal Data Protection*. 2005. Montreux, Switzerland.