ROTTERDAM UNIVERSITY OF APPLIED SCIENCES (CMI)

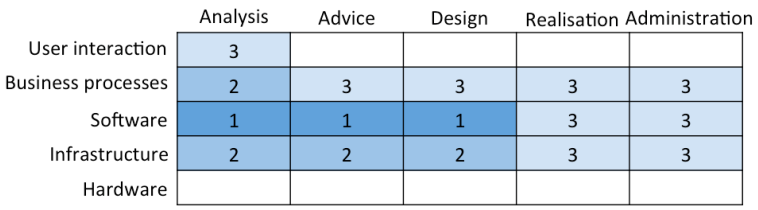
Privacy Engineering Methods (PEM)

For Enabling of Privacy by Design

CMIBOD01T

Number of study points:	3 ects
Course owners:	Tony Busker and Arne Padmos
Course guest lecturers:	Mortaza S. Bargh, Remon Cornelisse, Ronald Meijer and Sunil Choenni
Author(s):	Mortaza S. Bargh

Course overview

Module name:	Privacy Engineering Methods (PEM)
Module code:	CMIBOD01T
Study points and hours of effort:	This module gives 3 ects, in correspondence with 84 hours: • 8 x 3 hours frontal lectures • The rest is self-study
Examination:	Group assignment (with an oral presentation)
Course structure:	Lectures and self-study
Prerequisite knowledge:	Basics of data science (and knowing cryptography principles is encouraged)
Learning tools:	• Articles: At the end of each lecture a list of relevant articles will be given for further study • Lecture presentations (in pdf): To be found on N@tschool and GitHub repository `https://github.com/hogeschool/CMIBOD01T` • Homework/final assignments (in pdf): To be found on N@tschool and GitHub repository `https://github.com/hogeschool/CMIBOD01T`
Connected to competences:	
Learning objectives:	At the end of the course, the student has: • An understanding of privacy from a societal, legal, ethical, ... perspective • An understanding of and ability to apply some privacy enhancing technologies suitable for the data science domain (e.g., for data mining, data sharing and data storage) • An understanding about (a) making tradeoffs between data privacy and data utility and (b) choosing an appropriate set of privacy enhancing methods • Hands-on experience with some privacy-enhancing technologies
Notice:	Attendance in the lectures is compulsory
Course owners:	Tony Busker and Arne Padmos
Date:	November 15, 2016

1 General description

This section motivates the importance of the Privacy Engineering Methods course as part of data science education. The section also sketches the scope of such a course for ICT oriented educations within applied universities.

1.1 Why to consider privacy?

Nowadays one hears statements like "privacy is dead", "privacy is dying", "you have zero privacy anyway ... get over it"[1], "privacy might be a historical anomaly"[2], and "the age of privacy is over"[3]. Does privacy really vanish by the arrival and expansion of big data, open data, Internet of Things, social networks, data analytics, datacenters, ...? The answer to this question is NO! One should, indeed, pose another question here, namely: *What is meant by privacy*? If privacy is meant as the amount of information we can keep secret or unknown, then this privacy is perhaps shrinking. If privacy is meant as which rules should govern and how they should be applied to the production, collection, use and maintenance of personal information, then this privacy has never been more alive! As mentioned in [1]: "Privacy is very much alive, though it, like other social norms, is in a state of flux".

Protecting privacy has gained more importance in new laws and regulations, compared to that in the previous ones. Recently the European Commission's Safe Harbour Decision is invalidated by the European Court of Justice (on 6 Oct. 2015). The EU's new privacy protection law, called General Data Protection Regulation (GDPR), is going to be in effect from 2018 (expected). The GDPR will impose huge penalties (maximum of 4% of a company's/organisation's last worldwide annual turnover or 20M$, whichever higher) on those institutions that compromise privacy. Furthermore, commercial and public institutions embrace privacy protection not only as a legal obligation but also as a means of gaining the trust of customers and citizens or, in other words, they consider privacy protection as a means of distinction for their brands.

1.2 How to approach privacy?

Protecting privacy is a multi-disciplinary, inter-disciplinary and/or cross-disciplinary task. Privacy protection requires a close collaboration among multiple disciplines (such as ethical, judicial, technical, social, and psychological disciplines) from the very early stage on of a design process (i.e., adopting a privacy by design approach). This requires integrating non-technical and technical aspects, which in turn is possible if future data scientists gain enough education about the importance of and about the knowhow for approaching privacy by design from both non-technical and technical domains.

This course shall advocate a multi-disciplinary approach for addressing privacy and realising the privacy by design principles when designing socio-technical systems (i.e., those ICT systems used in their social context). Because the course scope is for ICT related educational disciplines, it will nevertheless have a technical focus (e.g., teaching more those methods and tools that support privacy preserving in storing, sharing, opening of privacy sensitive data).

1.3 Why to consider privacy in data science?

Success of data science applications depends on protection of privacy in the design, realisation, deployment and maintenance of such applications and systems. One important privacy threat occurring in data (intensive) applications is statistical disclosure, which refers to leaking (privacy) sensitive data when releasing/sharing micro-data or processed data. Therefore, it is necessary to carry out statistical data control before storing such data, sharing them with partner organisations, or opening them to the public.

This statistical data control steadily becomes difficult due to (i) growth of data volume, variety, velocity ... and (ii) availability of background information to data receivers and adversaries. The first factor makes it difficult to detect potential information privacy/sensitivity issues hidden in the released data. This factor relates to the intrinsic characteristics of the data, i.e., the privacy issues in a given dataset.

[1] Scott McNealy, Sun Microsystems CEO, Jan. 1999
[2] Vincent Cerf, Google Chief Internet Evangelist, 2013
[3] Mark Zuckerberg, Facebook, 2010

The second factor makes it difficult to assess the potential risks in combining the released data with other datasets (i.e., with background information). This second factor relates to the extrinsic characteristics of data, i.e., the privacy issues of a given dataset in relation to other datasets that exist in outside world due to, for example, sequential data release, multiple data release, continuous data release, collaborative data release, Big Data, and Open Data.

1.4 How to address privacy issues in data science?

One should augment human intelligence to do more than just applying some rules, laws and legislations in a simplistic way. This augmentation requires developing and using also the state of the art techniques, metrics and software tools for gaining insight into and remedying the potential intrinsic and extrinsic privacy (and information sensitivity) issues before, during and after data collection, storage and release.

Often privacy enhancing methods impact the quality (i.e., utility) of data. Therefore, one should take into account data utility after applying privacy enhancing techniques to the original data. This data utility/quality can be determined based on the purpose for which the data are collected, stored or released. There are metrics defined in literature for measuring data utility/quality. A fair comparison of different privacy preserving techniques, in other words, requires accounting for data utility after applying such techniques. This examination actually requires (i) measuring of data utility, (ii) measuring data privacy, and (iii) making the tradeoffs between data utility and data privacy in a given context in which the data are (going to be) processed. For example, the contextual constraints stem from organisational, business, judicial, ethical, etc. ideals and objectives. Note that measuring data privacy and data utility, and making tradeoffs between them, are prone to approximation and error, which should be taken into account appropriately when relying on such techniques.

1.5 Preliminaries

As preliminary knowledge for this course, the students are supposed to have some familiarity with data mining principles, cryptography principles, and basic mathematics.

2 Course program

The course is structured into 8 lectures. The lectures take place during eight (or seven) weeks of the term. The order according to which the lectures will be given may differ from the one mentioned below due to, for example, logistical or practical reasons.

2.1 Lecture 1 - Privacy foundations

The first lecture starts with giving an introduction to the the whole course. Subsequently, the lecture covers those non-technical part of the course that are relevant for understanding privacy in the context of data science.

Topics

- An introduction about the whole course (topics, house rules, end evaluation, etc.)
- Privacy definitions (a review of 6 definitions), viewing privacy based on the family resemblance theory
- Privacy threats, identified in the data lifecycle
- Privacy Impact Assessment (to be given by another guest expert)

Important references (and their denotations in the lecture presentations)

- Book: [2] denoted by [SOL'08]
- Articles: [3] denoted by [CRA'14]; [4] denoted by [CAV'12]; [5] denoted by [CAV'10]; [6] denoted by [HOE'14]

2.2 Lecture 2 - Privacy preserving data publishing: Non interactive

The second lecture is concerned with the basics of privacy protection when publishing/sharing structured micro data (e.g., tables in relational databases). It will give an overview of privacy threats and relevant privacy protection methods.

Topics

- Data anonymisation basics
- Privacy attacks: Record linkage, attribute linkage, table linkage, probabilistic attack
- Basic methods such as: Generalisation, suppression, anatomisation, permutation, perturbation
- Advanced solutions: k-Anonymity, l-Diversity, t-Closeness, etc.

Important references (and their denotations in the lecture presentations)

- Articles: [7] denoted by [FUN'10]; [8] denoted by [SWE'98]; [9] denoted by [MAC'07], [10] denoted by [LI'07]; [11] denoted by [XIA'06]; [12] denoted by [MAI'13]

2.3 Lecture 3 - Privacy preserving data publishing: Interactive

The third lecture is concerned with the basics of privacy protection when answering queries over micro data. It will give an overview of relevant privacy protection methods and review some privacy concepts from a technical perspective.

Topics

- Introduction to ϵ-differential privacy (for answering queries)
- Other applications of ϵ-differential privacy
- On the relation among identifiability, differential privacy and mutual-information privacy

Important references (and their denotations in the lecture presentations)

- Articles: [13] denoted by [DWO'06]; [14] denoted by [WAN'14]

2.4 Lecture 4 - Privacy preserving data publishing: Unstructured data

The fourth lecture is concerned with the basics of privacy protection when publishing/sharing unstructured data (e.g., transactional datasets as well as documents in natural languages). It will give an relevant privacy protection methods.

Topics

- Introduction to unstructured datasets
- Approaches for privacy protection in high dimensional datasets (optional)
- Approaches for privacy protection in unstructured datasets
- A practical application (to be given by another guest expert)

Important references (and their denotations in the lecture presentations)

- Articles: [15] denoted by [TER'08]; Liu2012c denoted by [LIU'12]; [16] denoted by [ZAK'14]; [17] denoted by [CAR'13]; [18] denoted by [GAR'09]

2.5 Lecture 5 - Privacy preserving data mining: Fundamentals

The fifth lecture is concerned with the basics of privacy protection when mining datasets. It will describe 4 or 5 data mining example methods in which privacy protection is the main concern. Furthermore, a taxonomy of privacy preserving data mining techniques will be given. Some metrics will be described that are proposed in literature to measure privacy in (un)structured datasets.

Topics

- A taxonomy of privacy preserving data mining methods
- Example cases I, II and III
- On privacy preserving data regression: Example case IV
- On privacy metrics

Important references (and their denotations in the lecture presentations)

- Articles: [19] denoted by [BER'05]; [20] denoted by [VER'04]; [7] denoted by [FUN'10]; [21] denoted by [ZLI'16]

2.6 Lecture 6 - Privacy preserving data mining: Cryptographic approaches

The sixth lecture provides an overview of Secure Multiparty Computation (SMC) principles and describes a number of main SMC methods at a high level. Although the described methods belong to the field of cryptography, there is no need to be deeply familiar with cryptography.

Topics

- SMC principles
- Homomorphic encryption: This part will cover those arithmetic circuits that allow certain algebraic operations to be carried out on the encrypted text (i.e., ciphertext).
- Garbled/Boolean circuits and their main building block called Oblivious Transfer (OT)
- Polymorphic Encryption and Pseudonymisation: This is a newly proposed method (mainly for medial datasets)
- Secrete sharing principles (optional, to be described in detail if there is time available).
- Fully-homomorphic: First plausible construction (by to be shortly mentioned).

Important references (and their denotations in the lecture presentations)

- Articles: [22] denoted by [KAR'07]; [23] denoted by [LIN'09]; [24] denoted by [PET'15]; Samet2015 denoted by [SAM'15]; [25] denoted by [SAM'08], [26] denoted by [SAM'09]

2.7 Lecture 7 - Other (partial) privacy enhancing techniques

The seventh lecture reviews a number of privacy enhancing technologies that cannot be categorised in the above-mentioned topics. The idea is to provide an overview of those mechanisms that aim at protecting privacy in other stages of data lifecycle than those mentioned so far. A subset of the topics mentioned below will be presented in this lecture.

Topics

- Access control and Usage CONtrol (UCON)
- Block chain for privacy protection applications
- Anonymous communication mechanisms: Mix networks (like Onion routing and TOR)
- Privacy in some practical cases (like smart meters, electronic voting, etc.)
- Use of high performance computing (optional)
- Privacy in identity management and anonymous credentials (optional)

Important references (and their denotations in the lecture presentations)

- Articles: [27] denoted by [CHO'16]; [28] denoted by [PAR'04]; [29] denoted by [HIL'05]; [30] denoted by [CLO'14]; [31] denoted by [ZUS'15]

2.8 Lecture 8 - From privacy by design to privacy engineering

The eighth lecture shall aim at coupling the high-level principles of privacy by design to privacy engineering techniques. Two systems will be also described to illustrate how to apply privacy by design principles in practice.

Topics

- Privacy protection guidelines
- Privacy by design principles
- Patterns of privacy methods
- Existing gaps and some guidelines to fill the gaps (i.e., putting puzzle pieces together)
- Describing two systems that are developed having the privacy by design principles in mind (to be given by another guest lecturer)

Important references (and their denotations in the lecture presentations)

- Articles: [3] denoted by [CRA'14]; [4] denoted by [CAV'12]; [5] denoted by [CAV'10]; [6] denoted by [HOE'14]; [32] denoted by [SCH'10]; [33] denoted by [CHO'10]

3 Assessment

For this course the students will be assessed based on

- Attending the lectures,
- Homework, and
- A group project.

3.1 Attending the lectures

Attending the lectures is compulsory. Because there is no classic material (like course book/booklet) available, it is important that students attend all sessions, where listening to lectures and having discussions are considered as main part of the learning proces. (Of course, there will be plenty articles that will be introduced from the literature for further study.) Attending at least 7 lectures will have positive impact on the score in case that there is a reasonable doubt about the quality of the final assignment.

3.2 Homework

At the end of lectures a homework assignment may be given, which must be handed over in written form by every student individually in the beginning of the next session. A couple of students will be chosen in the beginning of every session to share with the rest their answers to the homework of the previous session. *The lecturers still reserve the right to check the homework parts handed in by each student and to use it for further evaluation.*

3.3 Group project

From a set of possible projects, one project will be assigned to a group. Every group will consist of 3 students (in some exceptional situations groups of 2 persons are possible, with the approval of the main lecturer). The assignments will be given around week 4. The students may find/choose their own project, subjet to approval of the main lecturer. The group project is an assignment in the form of an applied research (desktop research, some hands on with software tools, analysis and experimentation, and conclusion).

3.4 Examination

For the evaluation purposes, every group should deliver a report (of 2500-3000 words) and a short presentation (of 10 minutes, thus 6-5 slides) three work-days before the final examination date (for example, if the latter is on a Friday, then the report and presentation should be given till end of the previous Tuesday).

The groups will present their results exactly the same way that they reported on the slides or the report three days before. Each group will have a short presentation of 10 minutes, answering a number of questions asked by other students or the lecturer(s).

3.5 Grading

The performance of every group shall be classified in three classes:

- FAIL (i.e., a grade 3)
- PASS (i.e., a grade 7)
- EXCELLENT (i.e., a grade 9)

The grading has three steps:

S_1 : If the final report and presentation are not delivered, the final grade is a FAIL (i.e., a grade 3), otherwise

S_2 : The oral presentation of the group will be assessed by the examinators. For this, the examinators shall use the assessment criteria (see below) and determine an outcome from the three classes mentioned above.

S_3 : If a grade is a FAIL and upon the request of the group, then the group's report and the homework parts will be used as extra evidences to support the grade given or possibly modify it.

3.6 Assessment criteria

The objectives of the course are determined as:

O_1 : Having an understanding of privacy from a societal, legal, ethical, ... perspective,

O_2 : Having an understanding of and ability to apply some privacy enhancing technologies suitable for the data science domain (e.g., for data mining, data sharing and data storage),

O_3 : Having an understanding about (a) making tradeoffs between data privacy and data utility and (b) choosing an appropriate set of privacy enhancing methods,

O_4 : Gaining hands-on experience with some privacy-enhancing technologies.

Based on the above mentioned objectives, we define the following assessment criteria:

C_1 : The research is carried out methodically (e.g., empirical, prototype based, ...). This means that the final report should contain problem statement (together with research questions), approach or methodology, literature review, algorithm or prototype (if needed), data/system evaluation, and conclusion.

C_2 : The group has adopted a multidisciplinary approach.

C_3 : The group has studied a (set of) privacy enhancing technology(ies) in depth.

C_4 : The group has considered the impact of the studied privacy enhancing technology on the data quality/utility.

C_5 : The group has considered the balance between data privacy and data utility aspects.

C_6 : The group has carried out enough experimentations with the data using the existing tools/algorithms.

The following table shows the relation between the course objectives and the evaluation criteria.

Criterion	Objective(s)
C_1 : Methodology and reporting	O_1, O_2, O_3, O_4
C_2 : Multidisciplinary approach	O_1
C_3 : Understanding technology	O_2
C_4 : Impact on data quality/usability	O_3
C_5 : Privacy and utility trade-off	O_3
C_6 : Empirical study	O_4

3.7 Grading, continue

For each of the above mentioned criteria, the group shall be classified at one of the three levels of L (low), I (Intermediary), and H (High). The final grade is determined as follows:

- FAIL (i.e., a grade 3): if the group gets a L for C_1 or two or more L in total,
- PASS (i.e., a grade 7): If the group gets at most one L for $C_2 \dots C_6$:.
- EXCELLENT (i.e., a grade 9): If the group gets for all criteria a H or if the group has added something new to the existing body of knowledge substantially.

The maximum grade counts.

3.8 Re-examination

According to the rules, the re-examination will be done in the following study period. For this, the same assignment should be carried out (unless otherwise mentioned). For this re-examination, there will be no support of the teaching staff.

Appendix: Learning matrix

Considering the Dublin-descriptors:

1. Knowledge and understanding
2. Applying knowledge and understanding
3. Making judgments
4. Communication
5. Learning skills

we can specify the course objectives accordingly, as summarised in the following table.

Learning goals	Dublin descriptors
O_1	1, 4 and 5
O_2	1, 2 and 5
O_3	3, 4 and 5
O_4	2 and 5

References

[1] N. M. Richards and J. H. King, "Big data and the future for privacy," *Handbook of Research on Digital Transformations (Elgar 2016)*, 2014.

[2] D. J. Solove, *Understanding Privacy.* Harvard University Press, 2008.

[3] K. Crawford and J. Schultz, "Big Data and due process - toward a framework to redress predictive privacy harms," *Boston College Law (BCL) Review*, vol. 55: 93, no. 1, 2014.

[4] A. Cavoukian and J. Jonas, "Privacy by design in the age of big data," *Information and Privacy Commisioner*, no. June, pp. 1–17, 2012.

[5] A. Cavoukian, "Privacy by design - the 7 foundational principles," *Identity in the Information Society*, vol. 3, no. 2, pp. 1–12, 2010.

[6] J.-H. Hoepman, "Privacy design strategies," in *ICT systems security and privacy protection*, pp. 446–459, Springer Berlin Heidelberg, 2014.

[7] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, 2010.

[8] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty*, vol. 10, no. 5, pp. 557–570, 2002.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.

[10] N. Li, T. Li, and S. Venkatasubramania, "t -closeness: Privacy beyond k-anonymity and l-diversity," *IEEE 23rd International Conference*, no. 3, pp. 106–115, 2007.

[11] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of the 32nd international conference on Very Large Database, ACM*, pp. 139–150, 2006.

[12] J. Maier, "Network simulation and its limitations," Tech. Rep. August, Technical University of Munchen, 2013.

[13] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, pp. 1–12, 2006.

[14] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy and mutual-information privacy," in *In 52nd IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1086–1092, 2014.

[15] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Anonymity in unstructured data," in *Proc. of International Conference on Very Large Data Bases (VLDB)*, 2008.

[16] H. Zakerzadeh, C. C. Aggrawal, and K. Barker, "Towards breaking the curse of dimensionality for high-dimensional privacy: An extended version," *arXiv preprint arXiv:1401.1174*, p. 13, 2014.

[17] D. Carrell, B. Malin, J. Aberdeen, S. Bayer, C. Clark, B. Wellner, and L. Hirschman, "Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text," *J Am Med Inform Assoc*, vol. 20, no. 2, pp. 342–348, 2013.

[18] J. Gardner and L. Xiong, "An integrated framework for de-identifying unstructured medical data," *Data and Knowledge Engineering*, vol. 68, no. 12, pp. 1441–1451, 2009.

[19] E. Bertino, I. N. Fovino, and L. P. Provenza, "A framework for evaluating privacy preserving data mining algorithms," *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 121–154, 2005.

[20] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM Sigmod Record*, vol. 33, no. 1, pp. 50–57, 2004.

[21] I. Žliobaitė and B. Custers, "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models," *Artificial Intelligence and Law*, pp. 1–19, 2016.

[22] A. F. Karr, W. J. Fulp, F. Vera, S. S. Young, X. Lin, and J. P. Reiter, "Secure, privacy-preserving analysis of distributed databases," *Technometrics*, vol. 49, no. 3, pp. 335–345, 2007.

[23] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy-preserving data mining," *Journal of Privacy and Confidentiality*, vol. 1, no. 1, pp. 59–98, Berkeley Electronic Press, 2009.

[24] M. Pettai and P. Laud, "Combining differential privacy and secure multiparty computation," in *Proceedings of the 31st ACM Annual Computer Security Applications Conference*, pp. 421–430, 2015.

[25] S. Samet and A. Miri, "Privacy preserving ID3 using gini index over horizontally partitioned data," *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, pp. 645–651, 2008.

[26] S. Samet and A. Miri, "Privacy-preserving bayesian network for horizontally partitioned data," *Proceedings - 12th IEEE International Conference on Computational Science and Engineering, CSE 2009*, vol. 3, pp. 9–16, 2009.

[27] S. Choenni, M. S. Bargh, C. Roepan, and R. Meijer, "Privacy and security in smart data collection by citizens," in *Smarter as the New Urban Agenda* (J. R. Gil-Garcia, T. A. Pardo, and T. Nam, eds.), ch. Volume 11, pp. 349–366, Springer, 2016.

[28] J. Park and R. Sandhu, "The UCON ABC usage control model," *ACM Transactions on Information and System ...*, vol. 7, no. 1, pp. 128–174, 2004.

[29] M. Hilty, D. Basin, and A. Pretschner, "On obligations," *Computer Security–ESORICS 2005*, pp. 98–117, 2005.

[30] P. Colombo and E. Ferrari, "Enforcing obligations within relational database management systems," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–14, 2014.

[31] G. Zyskind, O. Nathan, and A. S. Pentland, "Decentralizing privacy: Using blockchain to protect personal data," in *Security and Privacy Workshops (SPW)*, IEEE, 2015.

[32] Y. Schikhof, I. Mulder, and S. Choenni, "Who will watch (over) me? Humane monitoring in dementia care," *International Journal of Human Computer Studies*, vol. 68, no. 6, pp. 410–422, 2010.

[33] S. Choenni and E. Leertouwer, "Public safety mashups to support policy makers," in *Proceedings of the 1st International Conference on Electronic Government and the Information Systems Perspective (EGOVIS 2010)*, vol. 6267, (Bilbao, Spain), pp. 234–248, 2010.