Sentinel e-health network on grid: developments and challenges

Paul DE VLIEGER [a,b,c 1], Jean-Yves BOIRE [b], Vincent BRETON [b], Yannick LEGRE [d], David MANSET [d], Jérôme REVILLARD [d], David SARRAMIA [a,b] and Lydia MAIGNE [a,b]

[a] *Clermont Université, Université Blaise Pascal, LPC, BP 10448, F-63000 Clermont-Ferrand*

[b] *CNRS/IN2P3, UMR 6533, LPC, F-63177 Aubière*

[c] *ERIM, 28 Place Henri Dunant, BP38, 63001 Clermont-Ferrand cedex, France*

[d] *MAAT-France, 74070 Archamps, France*

Abstract. Grid technologies have proven their capabilities to settle challenging problems of medical data access. The grid ability to access distributed databases in a secure and reliable way while preserving data ownership opened new perspectives in medical data sharing and disease surveillance. This paper focuses on the implementation challenges of grid-powered sentinel networks within the e-sentinelle project. This initiative aims to create a lightweight grid dedicated to cancer data exchange and enable automatic disease surveillance according to definition of epidemiological alarms. Particularly, issues related to security, patient identification, databases integration, data representation and medical record linkage are discussed.

Keywords. Sentinel network, lightweight grid, cancer, surveillance, data linkage, epidemiology, patient identification

1. Introduction

In the cancer context, large amounts of medical data are being created, and the needs of communication and medical exchange growth. However, the current systems don't fulfil the requirements in terms of security, traceability, right management, patient privacy and also data ownership. Moreover, the centralised technologies currently used don't permit a fully connected system able to monitor the entire cancer activity over a geographic area. These solutions need several time consuming steps to deploy a set of data as no automatic methods exist from data extraction, representation, identification and currently everything has to be done by hand.

New opportunities offered by healthgrids, focused in [3], and in particular possibilities in terms of data sharing allow an easy integration of distributed medical datasets. Practically, if a grid can federate a large amount of distributed data, the capacity for large-scale statistics and epidemiology are vast.

This article reports an initiative to build a grid-based sentinel network able to federate heterogeneous and geographically distributed databases in a common

[1] Corresponding Author: Paul De Vlieger, vlieger@clermont.in2p3.fr

framework. This proposal opens new perspectives in data sharing, real-time monitoring and surveillance by giving a permanent link to data producers in oncology.

This article refers on current developments within the e-sentinelle project [1], [2] and focus on specific challenges to build lightweight grid-based sentinel networks, particularly medical data sharing, patient identification, data linkage, authorisation and security.

2. The e-sentinelle project

The e-sentinelle project [1] is a regional innovative initiative to build a grid-based sentinel network for cancer and precisely breast, colon and cervical cancers. These cancers are included in the screening campaigns of most of developed countries of UE [4]. The mortality impact of these screening structures has been proved through the world, in particular for breast cancer [5]. In France, the structures that have in charge of realising cancer screening have to invite a targeted population to be screened and have to ensure their follow-up.

In case of cancer detection, a biopsy is performed by surgery and analysed by a pathology laboratory. Then, a pathological report is created for the patient containing tumour analysis results (malignant/benign) with progression states, dimensions, etc... These reports are really relevant since they contain virtually complete data needed to monitor the cancer activity.

The screening follow-up requires a collect of these reports, but there are no electronic exchanges so data is printed and mailed by laboratories and recorded again by the screening structure. This process is costly and errors prone. In addition, data quality and thoroughness is not sufficient to produce reliable and comprehensive statistics on the impact and relevance of screening activity. Moreover, pathology labs can be reluctant to send their data since it is at their own expense (time, paper and postage). By the way, screening structures have to find information somewhere else, and most often besides the patient himself.

2.1. Goals

When the project started, two distinguished objectives raised:

- Enable a nominative data sharing between cancer fighting actors to improve collaborations, speed up treatments and ease follow-ups.
- Develop a regional structure to support large scale epidemiology analysis on data sources and allow monitoring or alerts specification for global health issues. Of course the second objective doesn't require nominative data, but only double free datasets with a correct patient identification and disambiguation.

The main requirement for data holders is to keep inside their place the data they produced. It appears that the grid technology is particularly well fitted to tackle this issue. A proof of concept has already been done with projects like eDiamond[6], ACGT[7] or Health-e-Child[9].

2.2. Architecture

The architecture of the project, as presented in [1], presents a typical sentinel network deployment. The central services (VOMS authentication [10], patient identification, etc...) are hosted in a neutral place, as no critical information (patient data) is stored inside these servers. Then, in each medical structure which takes part to the sentinel network, a grid node is deployed, with two interfaces:

- One linked to the medical database to expose, in the private medical structure network, this link can be database-view based, (e.g. a permanent and connected link to the production database) or by an automatic export system (SQL query on the database, and a standardized output). Even if the database link is not permanent, the second method is preferable. It is less intrusive, offers enlarged customisation for data integration and standardisation and does not overloads the production database during working hours. The export request can easily be scheduled during night.
- The second interface is linked to the sentinel network (Internet), enabling external users to query the grid data server, according to the local security and authentication policy fixed by VOMS.

For clients, the java client portal can easily be deployed on any workstation with an Internet access. All the business logic part of the application is transposed to a remote server according to the Pandora Gateway developed by the maat-Gknowledge company. This gateway is able to integrate various medical data according to openEHR specifications and expose them throughout the grid. The client application can launch requests on the sentinel network as large scale epidemiological studies.

2.3. Security, legal & ethical issues

The security and privacy requirements for a distributed medical data querying system are really important, and data protection is essential. Within the caBIG and ACGT projects, different studies about security, privacy, ethical and legal requirements have been published [8], [11]. The EU released a document [16] relative to personal data process, treatment and movement issues.

In this project the security is reinforced by using CPS healthcare professional smartcard [13], [14] released by the French health ministry, these cards will be available in the whole EU [15]. Practically, the chip contains an X509 grid-compatible certificate issued by a trusted certification authority. The authentication process and the data encryption are then ensured by these cards.

3. Management of patient medical data on a distributed architecture

3.1. Patient data consistency

A good consistency of distributed databases is fundamental. Several steps have to be followed before the integration of datasets in the grid sentinel network:

- Dataset standardisation: identification and organisation of the relevant fields, in partnership with health professionals.
- Data specification: semantic and metadata specification on fields.

- Data extractions: automatic patient data extraction mechanisms (scheduled or using views).
- Data integration: import patient datasets on a grid node to make them available to the sentinel network, adjusting rights according to local policy.

Regarding medical data, a patient identification step is needed to ensure a good level of coherency.

Some commons fields are required to allow an efficient record linkage. Minimal mandatory information is surname, first name, sex and birth date.

3.2. Patient identification

Throughout the health-care systems, cases of false patient identification are numerous and could be very responsible for mistakes in drug delivery or healthcares deliver to the patient. Due to a lack of global identification system, there is no solution to address the distributed patient identification issue. Most countries in EU have already a robust identification system (Refs). In France, the usage of the social security (SSN) number is strictly prohibited for data linkage as it contains privacy data: gender, date and place of birth… Moreover the accuracy and reliability of these numbers are reconsidered: the SSN number in US presents a high risk of identity. Aware of this issue, EU has launched the European Patients Smart Open Services (EPSOS) program [17] in order to build a European Electronic Health Record while the French government released guidelines to build a national health identifier [18]. Despite this, there is no suitable solution today; therefore a dedicated solution has been designed for this project. One of the strongest requirements of this identification system is being able to integrate third-party identification sources if a national solution is proposed in a near future.

As presented in [1], a new identifier is added to patients in each database. This identifier, uuid-based[19], is 128 bits long and can guarantee uniqueness across time and space (10^{38} possibilities). Moreover, it is fully anonymous as it is randomly generated (ex: `a5b4c706-d395-4c5e-8b44-62af9b11c43a`). However, due to his length and complexity, this number is not easily human readable and would be used only for data linkage.

The distributed identity management require specific ability to compare records and link identities. The entire reliability of the sentinel network depends on a good record linkage.

3.2.1. Identification system

The digital identity of a patient is naturally distributed and is independent for medical structure. As explained in **Figure 1**, each data provider has already a local patient number. During the data integration to the sentinel network, a new uuid number is allocated for each patient.

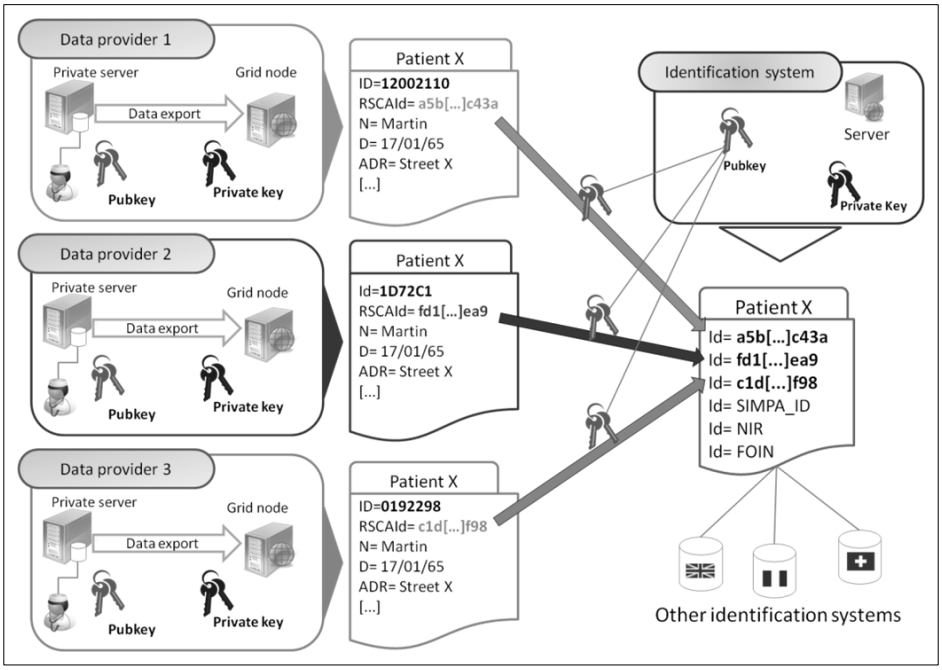

Figure 1. Identification system.

Then, the identification process consists in a matching between local and distributed identities using a central identity server. This server stores a list of identifiers for each patient. Using asymmetric cryptography of the Grid Security Infrastructure (GSI) included in the Globus project, the data and the generated uuid are always encrypted through the network, ensuring security and patient privacy. If the local patient matches another one in a remote database, the central server adds this uuid to the list of identifiers for a patient.

The management of multiple identifiers for a patient has several advantages:

- The scalability: the patient identity can evolve as easily as adding/removing a line to the central identification database.
- The privacy: as no common identifiers exist in distributed databases (no linkage possible without the central server) the patient privacy is ensured.
- Management of identities: in case of double identification for a patient or in case of two patients with the same identifier, merging or separating identities is straightforward.
- Interconnectivity: the connection to other identification sources consists simply in adding identifiers to the patient.

3.3. Data linkage

In order to link patient folders in a distributed way, we need to use a systematic comparison between available fields in the different patient's sheets. Although this method is purely practical it has proven his efficiency in different cases [20].

Epidemiological studies deserve reliable and uncorrelated patient data. Therefore the process used is there to link distributed identification numbers automatically. Ideally, the method must offer the best automatic linkage by restricting double-counting and avoiding false matching. **Figure 2** shows a theoretical situation where four areas set apart:

- True negatives for which no matching records is found in the database.
- True positives for which there are real matching between patients identification numbers.
- False negatives (or type I error), for matching failures (e.g. two identifiers for only one patient).
- False positives (or type II error) for wrong attribution of identity (e.g. two existing patients with one identity) this case must be minimized.

To avoid false positives the decision threshold must be gauged as accurately as possible (Ⓑ line).

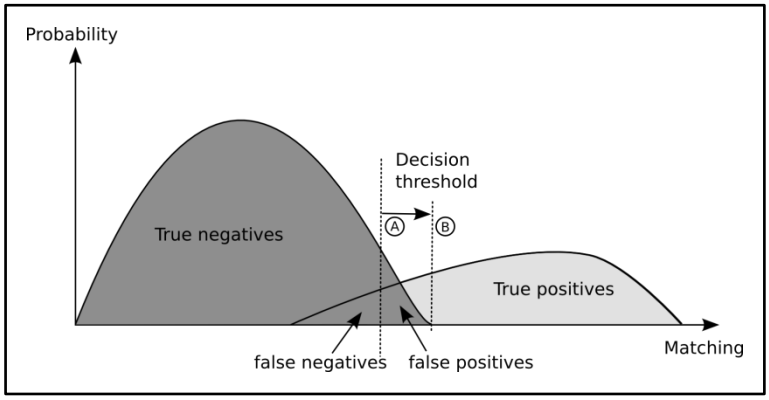

Figure 2. Theoretical matching problem; overlapping area is thin and a unique decision threshold is possible

Table 1 summarizes the different classification types and possible matching errors during the process.

Table 1. Classification and errors types.

Type	Reality	Classification
True positive	Match	Match
True negative	Non-Match	Non-Match
False negative	**Match**	**Non-Match**
False positive	**Non-Match**	**Match**

3.3.1. Solution proposed

Practically, as shown in **Figure 3** the false positives area is wider. If the threshold is kept as it, the linkage process will produce an unacceptable proportion of false positives. Thus, two solutions are possible:

- Adjust the Ⓑ threshold to an acceptable level, but the proportion of matching success will be really restricted to only perfect matching.

- Create other thresholds (① and ②) which define a new area between two automatic decisions (true positive and true negative). This new central area needs a manual intervention to determine if the patient matches or not.

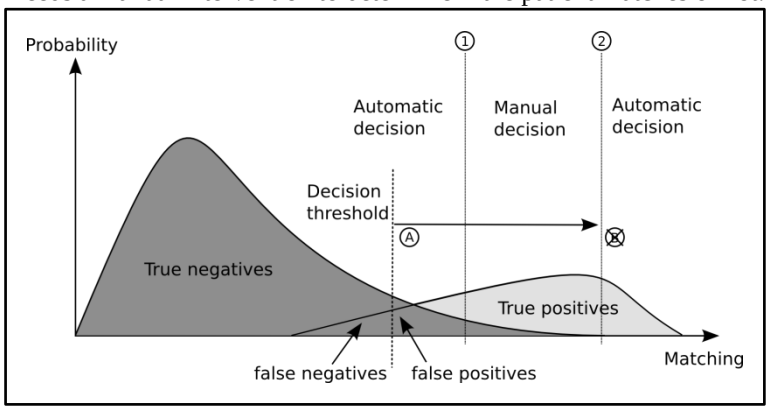

Figure 3. Practical matching problem; overlapping is wider and automatic decision isn't efficient, introduction of a second threshold for manual intervention.

The identification workflow for one database, as shown in **Figure 4**, consists in a standardisation of input data by identifying and normalising common fields to enable fusion-indexation of the dataset. Then, the most important step is an n-to-n systematic comparison between a patient and the whole distributed datasets of already integrated databases. A score is attributed (focused in section 3.3.3) and this input is processed by a classification algorithm which:

- Aggregates identities if `score>high_thres` (e.g. already known patient).
- If `score<low_thres` creates a new patient with a new identity (e.g. unknown patient).
- Asks for a manual intervention if `low_thres<score<high_thres` as automatic linkage present risks of errors.

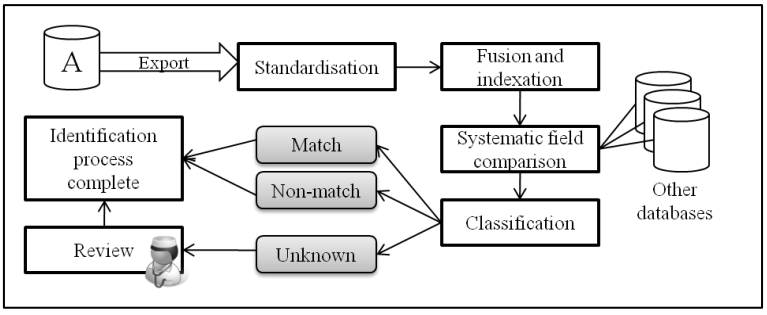

Figure 4. Matching process.

The manual intervention is performed by a dedicated person who decides to identify a patient (or not) by visualising data, with similarities and ambiguities highlighted by the system. If a client wants information for a specific patient, this system is also able to propose a set of possible matches and let the user choose the best correspondence.

As grids are naturally distributed, the risks database unavailability during the systematic comparison process is significant (server temporally down, authentication refused) so the automatic linkage process restarts when data becomes again available. The identification system was created to handle the grid flexibility so the possible bias introduced in the patient matching process can easily be handled by merging or separate identities.

3.3.2. Data linkage issues

Data linkage does not consist in a simple string comparison like `strcmp` in C, the two main problems are related to look alike patients (homonyms, same address, equivalent birthdates) and overall errors in names. According to [21] three levels of errors appears:

- Typographical errors (despite known spelling).
- Cognitive errors (comprehension problem).
- Phonetic errors (similar spelling).

The errors and variations are mainly related to typing of handwritten data, keyboard neighbours (k-i, e-r, etc...), input data during a telephonic conversation, software or database limitation of input fields (length limitation) that force the use of abbreviations or initials.

Several matching techniques exist and aim to measure similarity between strings. Two different approaches can be named:

- pattern matching, for flexible matching between two string.
- a combination of phonetic encoding and exact matching.

The similarity measure is generally normalised: if two strings are equivalents the score=1 and if totally different score=0.

3.3.3. String similarity methods

Pattern matching measures distance between two strings. One of the most powerful string similarity comparison methods is the Jaro-Winkler algorithm [22]. The mechanism is simple and consists in a score calculation based on the sum of exact characters matches and permutations. An improved, but slower version, Permuted-Winkler, calculates Winkler score for all words permutations and returns the maximum. This method is really efficient for full names as given names/surnames permutations are numerous… Another efficient method, Longest Common Sub-string, compares similarity between words by removing the longest common part of the two strings.

3.3.4. Phonetic-based methods

The other class of matching techniques are Phonetic-based. The first methodology, Soundex, defines a numerical equivalent to English sounds [23] (**Table 2**) and applies a phonetic transformation to words according to this transformation table, missing letters are insignificant.

Table 2. Soundex English transformation table.

Class	**1**	**2**	**3**	**4**	**5**	**6**
Letters	B F P V	C G J K Q S X Z	D T	L	M N	R

The codification consists in keeping the first letter and adding the first three digits of the Sounded transformation. For example, Jack=J200, Jacob = J210, Jacobsen = J212, Catherine= Catarina = Catarinella = C365 (but Citron=C365). Soundex is also highly pronunciation-dependant and adaptations are needed to be efficient on different languages, different alphabets and pronunciations.

Phonex is an improved version of Soundex which requires a language-sensitive string pre-processing before codification. More the Phonex pre-processing have a thorough knowledge of the language pronunciations, more the output code is accurate. The number of transformations can be time-consuming and slow the process if a lot of comparisons are requested. To speed-up the process, a codification cache system can be implemented for a set of commons strings like most used surname, etc... Moreover, the process can be easily parallelised as comparisons are completely independents.

3.3.5. Solution adopted for the e-health Sentinel network

The proposition of solution must be as efficient as possible to maximize the number of automatic matching. For this patient linkage process the usage of a combination of Jaro-Winkler and Phonex (French) algorithms is used. According to the relevance and accuracy of information in the dataset, different weights are attributed.

For each field, four different criteria define how to interpret matching scores according to field types:

- Accuracy, which defines the relevance of information.
- Blocking, in case of false matching (under threshold), the correspondence would be automatically rejected.
- Weight (similar), which represents a factor attributed in case of similarity (over threshold).
- Weight (different), in case of false matching, a divide factor attributed to global similarity.

Weight distinction between similar and different matching is necessarily as for example: probability to have a different last name for only one patient in distributed databases is small so it reduces considerably the matching chance. However, two entries with the same address don't mean that the patient is the same for these two entries. **Table 3** summarises the proposition of criteria adjustment for automatic record linkage. A weight factor is attributed for each field and is submitted as input for the linkage algorithm.

A global score is attributed for each n-to-n comparison and is submitted as input for the matching process.

Table 3. Relevance of information for selected fields.

	Last name	**First name**	**Sex**	**Maiden name**	**Birth**	**Address**	**Region**	**Postcode**	**City**	**Physician**
Type	String	String	Digit	String	Date	String	Digit	Digit	String	String
Accuracy	•••	•••	•••	•	•••	•	•	••	•	•
Blocking	X		X		X					
Weight (similar)	•••	••• *	••*	•	•••	••	•	••	••	••
Weight (different)	•••	• *	••	••	••	•	••	•	•	•

* Only if previous fields matches

3.4. Security

Security is a really important issue. Patient data contain information which can easily identify the owner, either directly (name/surname) or using non-anonymous identifiers (which contains pieces of patient data). The medical context of this project implies the highest consideration on security and privacy issues. The linkage of distributed databases bound to a single individual presents many privacy risks and must be handled by the system. The linkage process guarantee that only the identity server is able to join identities and this server log each request. The use of smartcard certificates for the authentication process is responsibility of the owner.

Since the beginning of grid computing, the developers' community always takes into account security measures. Authentication and authorisation are fully integrated and proved his hardiness, but they are still basic security mechanisms, and the business logic part of the network (web services) must strengthen security.

4. Data linkage results

In this section we discuss experiments using different data linkage algorithms described previously to identify strengths and weaknesses of these methods and in order to adjust different thresholds.

Using the US Social Security Death Index, publicly available at [24] we retrieved two subsets of 10000 unique records with 12% of patients present in both datasets. We selected only 4 fields: first/last name, birth date and address (last residence). Without any modification, the matching process is successful, only perfect homonyms causes troubles and needed manual intervention (only birth dates or address were different).

In order to estimate the hardiness of the different methods, we integrated random bias in the datasets, using character inversions or replacements with keyboard neighbours or random characters.

The preliminary results, as shown in **table 4**, use a threshold of 0.8 for automatic linkage, the different columns represent true positive (TP), false negative (FN), false positive (FP) , result (TP/12), and the accuracy is calculated using TP/(TP+FN+FP).

Table 4. Results, in matching percentage over a 10000 entries dataset with 12% of joint patients

Field – Method	TP	FN	FP	Result	Accuracy
Last Name – Jaro-Winkler	11.53	1.21	0.06	96.08	90.08
Last Name – Soundex-US	9.33	1.14	0.11	77.75	88.19
First Name – Jaro-Winkler	13.11	2.21	0.09	109.25	85.07
First Name – Soundex-US	10.37	1.93	0.13	86.42	83.43
Address – Jaro-Winkler	9.82	1.72	0.11	81.83	84.29
Address – Soundex-US	7.41	1.72	0.19	61.75	79.51

Globally, the Jaro-Winkler method has better results concerning matching score but the percentage of false negative and false positive are significant (at least 121 false negatives and 6 false positives over 10000 records). The Jaro-Winkler marching result for first name exceed 100% because of people having the same first name. In the real matching process, the result of both first and last name is taken into account.

Since only one field isn't sufficient to correctly link datasets, the combination of matching results of a maximum common fields and different matching methods on the dataset is needed.

These preliminary results gave us an overview of data linkage possibilities and have to be enhanced and tested on real case datasets with real errors in the field description. These datasets will be soon available when the production phase of the e-sentinelle project will start (scheduled in mid-2010). The different thresholds will have to be precisely defined to maximise the automatic process efficiency.

5. Conclusion

The e-sentinelle initiative, started in 2009, aims to build a grid-enabled sentinel network on cancer and demonstrates the relevance of grid technologies to address disease surveillance. The development progresses of the project highlighted some specific issues and this paper reports our efforts to settle some of them. One of the most important problem was how to manage a patient identity over a distributed network. Indeed, most of data owners want to keep control on their data, so no complete extraction is possible. As grids can easily federate distributed databases, this major requirement was met. But patient identity is then distributed and the goal is now to link this identity among the grid network. By using distributed identification mechanism in combination with data linkage techniques, patient matching problem is practically settled. Only a validation step on real datasets has to be done.

Acknowledgements

The work described in this article was partly supported by grants from the European Commission EGEE, Embrace), the French Ministry of Research (GWENDIA) and the regional authorities (Conseil Régional d'Auvergne, Conseil Général du Puy-de-Dôme, Conseil Général de l'Allier).

The Enabling Grids for E-sciencE (EGEE) project is co-funded by the European Commission under contract INFSO-RI-031688. The EMBRACE project is co-funded by the European Commission under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092. Auvergrid is a project funded by the Conseil Regional d'Auvergne. The GWENDIA project is supported by the French ministry of Research.

References

[1] P. De Vlieger et al, Grid-enabled sentinel network for cancer surveillance, *Studies in health technologies and informatics* **147** (2009), 289-94.
[2] Sentinelle project, www.e-sentinelle.org
[3] V. Breton et al, The Healthgrid White Paper, *Studies in Health Technology and Informatics*, **112** (2005), 249-321.
[4] Cancer Screening in the European Union; Report on the implementation of the Council Recommendation on cancer screening. 2007.
[5] T. Morimoto et al, Current status of breast cancer screening in the world, *Breast Cancer* **16** (2008), 2-9.
[6] M. Brady et al, eDiamond: a grid-enabled federated database of annotated mammograms, *Grid Computing: Making the Global Infrastructure a Reality F Berman*, G Fox and T Hey (eds), Wiley, 2003.

[7] ACGT project, http://www.eu-acgt.org/
[8] ACGT, legal and ethical requirements: http://www.eu-acgt.org/documents/public-deliverables.html
[9] The Health-e-Child project: http://www.health-e-child.org/
[10] R. Alfieri, R. Cecchini, et al, From gridmap-file to VOMS: managing authorization in a Grid environment, *Future Generation Computer Systems* **21** (4) (2005): 549-558.
[11] FJ. Manion et al, Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study, *BMC Medical Informatics and Decision Making* ***9*** (2009) 31.
[12] OpenEHT, http://www.openehr.org
[13] GIP-CPS, http://gip-cps.fr
[14] P.Fortuit, Professional health cards (CPS): informatic health care system in France, *Ann Pharm Fr* **63** (5) (2005), 350-5.
[15] HPRO Card, http://www.hprocard.eu/
[16] Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, L281/31 23.11 (1995).
[17] http://www.epsos.eu/
[18] http://www.asipsante.fr/
[19] P. Leach, M. Mealling, and R. Salz. A Universally Unique IDentifier (UUID) URN Namespace. IETF RFC 4122 (2005), http://www.ietf.org/rfc/rfc4122.txt
[20] P. Christen, A comparison of Personal Name Matching: Techniques and Practical Issues, Technical report, TR-CS-06-02, Computer Science Laboratory, The Australian National University, Canberra, Australia.
[21] K. Kukich, Techniques for automatically correcting words in text, *ACM Computing Surveys* **24**(4) (1992) 377-439.
[22] WE. Winkler, Overview of record linkage and current research directions, *Technical Report RR2006/02, US Bureau of the Census* (2004).
[23] RC. Russell US Patent 1435663 (1922).
[24] Social Security Death Index, http://ssdi.rootsweb.ancestry.com