GINSENG, une infrastructure de grille au service de l'e-santé et de l'épidémiologie

Sébastien Cipière[1,2,3], Sébastien Gaspard[5], David Manset[5], Jêrome Revillard[5], David Sarramia[1,2], Vincent Breton[1,2], David R.C. Hill([1,2,4], Lydia Maigne[1,2]
cipiere@clermont.in2p3.fr, sgaspard@maatg.fr, dmanset@maatg.fr, jrevillard@maatg.fr, sarramia@clermont.in2p3.fr, breton@clermont.in2p3.fr, david.hill@univ-bpclermont.fr, maigne@clermont.in2p3.fr.
(1) Clermont Université, Université Blaise Pascal, LPC, BP 10448, F-63000
(2) CNRS/IN2P3, UMR6533, LPC, F-63177
(3) CNRS, UMR 6158, Université Blaise Pascal, LIMOS, F-63173
(4) ISIMA, Institut Supérieur d'Informatique, de modélisation et de leurs Applications, BP 10125, F-63177
(5) MAAT France, GNUBILA Group, Argonay, F-74370

Overview:

The GINSENG (Global Initiative for Sentinel E-health Network on Grid) project aims to implement a grid infrastructure for e-health and epidemiology in Auvergne. A distributed medical database is created upon a secure network for epidemiological studies. Our goal is to create a decentralized information system using grid technologies. The medical sites involved in the project are clustered around two themes: cancer monitoring and perinatal care. On each medical site a server which duplicates the medical database, is deployed with grid services. At the same time, full control of the information is kept by the organizations storing patients' files. This solution allows for a high level of security, privacy, availability, and fault tolerance. Queries made on the distributed medical databases are made via a secure web portal. Public health authorities use this infrastructure for health monitoring, epidemiological studies and evaluation of specific medical practices.

Mots clés: grille, bases de données, e-santé, épidémiologie, réseau de surveillance.

Enjeux scientifiques, besoin en calcul, stockage et visualisation :

Bases de données médicales distribuées

Le projet GINSENG vise à mettre en réseau des bases de données médicales existantes et complémentaires dans le but de compléter les informations médicales manquantes à tout dossier de patient suivi dans le cadre d'une pathologie. Les applications médicales choisies pour illustrer le principe de partage de la donnée médicale sont le suivi des cancers dans le cadre du dépistage organisé et le suivi des parturientes.

Les bases de données médicales peuvent être de nature très hétérogène, elles comprennent des informations médicales textuelles ainsi que des imageries mais dans des domaines médicaux très différents comme par exemple une analyse de prélèvement sanguin (au format texte) qui pourra compléter une information issue de l'imagerie comme la clarté nucale chez le fœtus potentiellement atteint de trisomie 21. Ces données médicales ne peuvent avoir un sens pour une analyse pertinente en santé publique si elles répondent à certains besoins des épidémiologistes :

- L'information médicale doit être fiable, les données médicales doivent être chaînées de manière performante pour une compréhension globale de la pathologie médicale.
- L'information doit être accessible en temps réel par tout utilisateur du réseau étant autorisé
- L'information médicale doit être exhaustive ce qui implique qu'elle puisse être complétée en temps réel également

Pour répondre à ces enjeux, les bases médicales doivent être interconnectées efficacement en utilisant un format de données standardisé. Le choix du type de format standardisé doit être fait de manière à s'adapter à toute structuration déjà établie de bases de données.

Dans le cadre de ce projet, les données médicales restent pour la plupart textuelles avec néanmoins une utilisation d'imagerie médicale peu volumineuse (mammographie ou échographie dont la taille n'excède pas quelques mégaoctets.).

Identification et chaînage des données médicales

Le système a pour but de mettre en commun des bases de données provenant de divers sites médicaux. Il est alors primordial de s'assurer de l'identité des patients lors de l'utilisation de leurs données médicales pour des études épidémiologiques. Il est possible par exemple, qu'un patient consulte dans plusieurs sites médicaux, d'autre part la probabilité qu'un homonyme fréquente le même établissement de santé n'est pas négligeable. Il est aussi envisageable que lors de l'enregistrement d'un dossier dans le système d'information une erreur se soit produite. Ces trois différents cas doivent être correctement pris en charge par l'infrastructure. En effet en aucun cas le système ne devra associer des patients distincts à une même identité. De la même façon qu'il est important que le système puisse associer deux dossiers provenant de deux services d'information différents au bon patient. Comme notre solution est développée sur le territoire français elle respecte les critères de la Commission Nationale de l'Informatique et des Libertés (CNIL) qui garantit notamment le respect des conditions relatives à la gestion des bases de données intégrant des informations personnelles. Il s'agit donc d'utiliser des algorithmes d'identification performants et rapides pour chaîner l'information médicale d'un même patient sur différents sites hospitaliers.

Sécurité
À partir du moment où une information médicale est utilisée par une tierce personne ou bien transite via un réseau, les mécanismes de sécurité afférents doivent être très aboutis. La sécurité doit être présente à différents niveaux :

- au moment de l'authentification d'un utilisateur sur le réseau. Celui-ci doit être reconnu et ses droits d'accès doivent être clairement notifiés.
- Lors du stockage des informations médicales dans les bases données. Ce stockage doit être réalisé de manière encryptée.
- Lors du transit de l'information sur le réseau, le partage de la donnée médicale devra lui aussi être encrypté.
- Le réseau mis en place devra lui aussi être sécurisé et compartimenté pour chaque application du projet.

Ces contraintes de sécurité doivent répondre également à toutes les contraintes légales françaises et européennes. Pour la France, la création de la « Loi relative à l'informatique, aux fichiers et aux libertés du 6 janvier 1978 ». Cette loi encadre le traitement de l'information en France, en stipulant que l'informatique « Ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques ». Cette loi instaure aussi des droits aux citoyens, en leur donnant accès (opposition/rectification) aux données les concernant.
Dans la lignée de la loi Informatique et Libertés, l'Union Européenne s'est dotée, en 1995 d'un texte commun pour l'ensemble des pays membres: la directive Européenne 95/46/CE. Cette loi est largement inspirée du texte français mais est moins restrictive : elle n'impose pas une organisation de contrôle (indépendante) comme la CNIL.

Accès et visualisation des données
Une fois les données structurées et mises à disposition de manière sécurisée, il faut les rendre disponibles via une interface ergonomique et sécurisée. Les besoins des personnels médicaux, notamment en santé publique sont de 2 natures :

- Interroger les bases de données médicales de façon transparente et sécurisée.
- Visualiser l'information médicale agglomérée par des graphiques ou bien par téléchargement de fichiers (au format csv) pour un post-traitement avec des logiciels d'analyses statistiques comme SAS ou R (Bates et al., 2010).

Dans la partie suivante, nous reprenons chacun des points dont les enjeux ont été détaillés pour en expliciter leur développement.

Développements, utilisation des infrastructures :

Bases de données médicales distribuées
Chaque base de données médicale est dupliquée sur un serveur de grille appelé gateway. Les services présents sur cette gateway sont :

- Les services de grille standard glite BDII, LFC, WMS, LB et VOMS pour la glue du système.
- AMGA pour permettre la gestion de bases de données relationnelles distribués et hétérogènes.

La gateway est la partie visible de chaque site, toutes les gateway sont interconnectées pour former un réseau. Le réseau ainsi créé constitue notre base de données distribuée. Dans un souci de sécurité maximale et de disponibilité des données patients, nous pouvons mettre en œuvre différents protocoles tel que la fragmentation de la base de données et une duplication de ces fragments. Ce qui nous permet de continuer d'avoir accès à la totalité des informations contenues dans la base de données bien que certains sites puissent être inaccessible.
La structuration des données est réalisée sous le format FedEHR qui vise à fédérer les informations médicales sous une arborescence dans laquelle le patient est la racine de toute l'information médicale basé sur les principes énoncé dans (Benson, 2010) . Cette information est ensuite regroupée sous la forme d'évènements médicaux, eux-mêmes agrégeant des variables cliniques. Toutes les bases de données médicales auxquelles nous avons pu avoir accès à ce jour ont été interprétées sous la structure FedEHR (notamment les bases de données gynécologiques, de dépistage des cancers et d'anatomocytopathologie).

Identification et chaînage des données médicales
Pour répondre aux différentes problématiques liées à l'identification des patients nous avons retenu deux algorithmes : Jaro Winkler (Jaro, 1995) et Soundex (Quantin et al., 2004) . L'algorithme Jaro Winkler est basé sur l'analyse de chaînes de caractères. L'algorithme Soundex est quant à lui un algorithme qui compare les données en se basant sur les phonèmes, ce qui permet de considérer « Philippe » et une version mal orthographiée de ce prénom telle que « filipe » comme étant similaire. Ces deux algorithmes auront à comparer les données des patients ; les informations que nous avons retenues sont le prénom, le nom, le sexe, la date de naissance, et l'adresse. Pour améliorer la vitesse de traitement des algorithmes nous étudions actuellement des solutions basées sur la puissance des cartes graphiques GP GPU de type Fermi (Owens et al., 2008), qui pourront éventuellement être déployées sur le serveur ayant la charge de l'identification des patients.

Sécurité
En raison du caractère sensible des données manipulées, le projet GINSENG dispose d'un réseau qui lui est propre. Ce réseau se situe pour sa plus grande partie au sein d'un Virtual Private Network (VPN) (IPSEC) qui assure une sécurité

accrue des données. Ce VPN est créé au travers de l'Internet et géré par des routeurs de marque Cisco. L'accès au service s'effectue par une page internet sur le site e-ginseng.org, pour qu'une connexion s'établisse l'utilisateur doit disposer d'une Carte de Professionnel de Santé (CPS). Cette carte est la clef d'accès au Serveur Central d'authentification (CAS). Le serveur central d'authentification s'appuie sur la technologie Virtual Organization Membership Service (VOMS), et la base de données distribuée utilise Arda Metadata catalogue (AMGA) (Koblitz, Santos, & Pose, 2007). Les données qui transitent sur le réseau sont encryptées bien qu'elles transitent exclusivement par le VPN pour encore plus de sécurité.

Accès et visualisation des données

Un serveur Web héberge le site Internet du projet, qui est le point d'accès unique à l'infrastructure. Le site Internet poursuit trois objectifs principaux, l'information, l'identification, et la gestion des requêtes épidémiologiques. Au travers d'une interface ergonomique et attractive, des informations statistiques sont rendues disponibles aux différentes catégories d'utilisateurs. Les utilisateurs sont identifiés par un serveur Central Authentification Service (CAS) grâce à leur carte de professionnel de santé (CPS). Le serveur CAS permet une seule et unique authentification de l'utilisateur, tout en lui octroyant l'accès à tous les services, ainsi c'est le serveur CAS qui authentifie utilisateurs auprès du Virtual Organization Membership Service (VOMS) (Alfieri et al., 2005). En fonction de leur profession et de leur rôle au sein de l'organisation, les personnes authentifiées n'ont pas accès aux mêmes tableaux de bord. Une Virtual Organisation (VO) (Cummings, Finholt, Foster, Kesselman, & Lawrence, 2008) ou Organisation Virtuelle a été créée spécifiquement pour les besoins du projet. Elle a pour nom VO Sentinelle.

Les requêtes effectuées sur les bases de données sont réalisées en langage SQL. Ces requêtes sont créées directement depuis le portail web e-ginseng.org et interprétées par toutes les gateways présentes sur les sites distants.

Résultats scientifiques :

Le projet GINSENG est en phase de développement. Les bases de données médicales utilisées pour tester les solutions techniques mentionnées ci-dessus sont pour l'instant simulées en adoptant une structuration des données identiques aux bases de données médicales réelles.

Identification et chaînage des données

À ce jour, la phase de développement a permis de valider les algorithmes d'identification Jaro Winkler et Soundex portées à la fois en langage C++ sur CPU (2.4 Ghz) et en langage CUDA sur GP-GPU de type Fermi (Tesla C2050).

Il a été montré que l'algorithme Jaro Winkler était plus rapide d'un facteur 14 à la comparaison de 5 millions de noms et prénoms. Une prochaine étape est de tester les seuils d'acceptance de comparaisons des 2 algorithmes en introduisant dans les bases de données simulées des erreurs comme l'nsertion de prénoms composés, l'inversion du nom et du prénom, l'ajout d'accents, le doublement d'une lettre, la suppression d'une lettre, l'inversion de 2 lettres ou encore le remplacement d'une lettre. Car une étude menée par (Friedman & Sideli, 1992) a montré jusqu'à 27% d'erreurs d'identification de patients au sein de trois bases de données hospitalières de 100.000 patients sur le même site.

Les applications médicales

Différents cas d'études épidémiologiques ont fait l'objet des premiers tests du réseau GINSENG. L'intérêt est de montrer comment un système tel que GINSENG peut permettre d'améliorer les études épidémiologiques en chaînant correctement les informations médicales provenant de sites médicaux différents. L'étendue des potentialités de GINSENG est démontrée par des études à grande échelle concernant :

- L'étude de l'incidence d'une pathologie
  - Exemple : Nombre de nouveaux cancers du sein sur la dernière année pour une ville donnée.
- L'étude des facteurs de risques d'une pathologie
  - Exemple : Recherche d'éléments influençant la survenue de cancers.
- La validation d'une information de nature déclarative pour optimiser la prise de décisions médicales
  - Exemple : croisement des informations communiquées par la patientes et les BDD disponibles.
- L'étude des pratiques médicales
  - Exemple : comparaison du taux de césariennes en fonction des maternités.

Perspectives :

Le projet est actuellement en phase de développement et de nouvelles solutions apparaissent régulièrement pour toujours plus sécuriser les données des patients et s'assurer de leurs validités. Il sera envisagé d'associer automatiquement les champs des bases métiers, aux champs FedHER de la base GINSENG au moyen d'une ontologie comme le propose (Faucher, Bertrand, & Lafaye, 2008). À moyen terme les 11 plus importants hôpitaux auvergnats et les principaux cabinets d'anatomocytopathologie seront équipés avec un serveur de type gateway pour prendre part au projet GINSENG. À long terme le projet permettra les l'échange des données entre les divers partenaires, ainsi il sera envisageable d'équiper des partenaires institutionnel tel que l'InVS (Institut de Veille Sanitaire), qui pourra actualiser ses données quotidiennement.

Les procédés que nous appliquons à la périnatalité et au suivi des cancers sont facilement transposables aux

problématiques qui pourraient se poser dans d'autres domaines, ou secteurs d'activité. Pour conclure un élargissement géographique à l'échelle de la France est également envisagé.

Références :

Alfieri, R., Cecchini, R., Ciaschini, V., dell' Agnello, L., Frohner, Á., Lorentey, K., & Spataro, F. (2005). From gridmap-file to VOMS: managing authorization in a Grid environment. *Future Generation Computer Systems*, *21*(4), 549-558. Elsevier. doi:10.1016/j.future.2004.10.006

Bates, D., Chambers, J., Dalgaard, P., Falcon, S., Gentleman, R., Hornik, K., Iacus, S., et al. (2010). The R Project. *Power*. Retrieved from http://www.r-project.org/

Benson, T. (2010). The HL7 V3 RIM. *Principles of Health Interoperability HL7 and SNOMED* (pp. 1-20). Springer London. doi:10.1007/978-1-84882-803-2_7

Cummings, J., Finholt, T., Foster, I., Kesselman, C., & Lawrence, K. A. (2008). *Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations*. *Technology* (Vol. 3, p. 52). National Science Foundation. Retrieved from http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

Faucher, C., Bertrand, F., & Lafaye, J.-Y. (2008). Génération d'ontologie à partir d'un modèle métier UML annot. *Revue des Nouvelles Technologies de IInformation RNTI*, *E*(12), 65-84. Retrieved from http://hal.archives-ouvertes.fr/hal-00382954/en/

Friedman, C., & Sideli, R. (1992). Tolerating spelling errors during patient validation. *Computers and biomedical research an international journal*, *25*(5), 486-509.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, *14*(5-7), 491-498. doi:10.1002/sim.4780140510

Koblitz, B., Santos, N., & Pose, V. (2007). The AMGA Metadata Service. *Journal of Grid Computing*, *6*(1), 61-76. doi:10.1007/s10723-007-9084-6

Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., & Phillips, J. C. (2008). GPU Computing. *Proceedings of the IEEE*, *96*(5), 879-899. ACM Press. doi:10.1109/JPROC.2008.917757

Quantin, C., Binquet, C., Bourquard, K., Allaert, F., Gouyon, B., Ferdynus, C., Pattisina, R., et al. (2004). Estimation de la valeur discriminante des traits d'identification utilisés pour le rapprochement des données d'un patient. *Revue d'Épidémiologie et de Santé Publique*, *52*(5), 431-440. doi:10.1016/S0398-7620(04)99079-7