Partage de données biomédicales sur le web sémantique

Rémy Choquet[1], Douglas Teodoro[3], Giovanni Mels[2], Ariane Assele[1], Emilie Pasche[3], Patrick Ruch[3], Christian Lovis[3], Marie-Christine Jaulent[1]

[1] INSERM UMRS872 EQ.20, Université Pierre et Marie Curie, 75006 Paris
{remy.choquet, marie-christine.jaulent, ariane.assele}@crc.jussieu.fr
[2]AGFA Healthcare, Ghent, Belgium
giovanni.mels@agfa.com
[3] SIM, Université de Genève et Hôpitaux Universitaires de Genève, Suisse
{douglas.teodoro, patrick.ruch, emilie.pasche}@sim.hcuge.ch

Résumé : L'explosion de la quantité de données à traiter et à partager, particulièrement dans le domaine biomédical, pousse la communauté à construire des systèmes intégrés où sémantique et données sont couplées. Le projet Européen DebugIT adopte les technologies du web sémantique pour partager des données liées à l'émergence de la résistance aux antibiotiques en Europe. Nous proposons une mise en pratique de ces technologies à travers un cadre d'interopérabilité technique, syntaxique et sémantique. Nous validerons l'approche sur des données réelles, multilingues et multi terminologiques.

Mots-clés : Intégration de données, interopérabilité sémantique, web sémantique.

1 Introduction

Le coût de stockage de l'information étant toujours plus réduit, nous avons connu au cours des dix dernières années une explosion de la volumétrie des données biomédicales disponibles (Galperin, 2008). Les bases de données couvrent aujourd'hui une part de plus en plus importante de l'information biomédicale : les données administratives des patients, les examens biologiques, les diagnostiques cliniques, les images, ou bien encore les données génétiques. Cependant, l'utilisation secondaire de cette masse d'information afin d'améliorer le soin et la sécurité du patient est encore limitée. Le développement d'un système qui puisse intégrer des données biomédicales à travers différents pays pose plusieurs problématiques : le manque de standards techniques (Sheth & Larson, 1990) ; la diversité de la sémantique des sources de données (Karasavvas *et al.*, 2004) ; la gestion de la qualité de données (Choquet *et al.*, 2010) ; et enfin la sécurité et la confidentialité des données patients qui doivent être préservées (Iavindrasana *et al.*, 2007).

La littérature propose trois approches afin de répondre à une partie des problématiques soulevées ci-dessus : l'approche entrepôt de données comme dans les projets BioWarehouse (Lee *et al.*, 2006) et BioDWH (Töpel *et al.*, 2008) ; l'approche de médiation (ou d'intégration par vues) dans les projets HEMSYS (Pillai *et al.,* 1987) et TAMBIS (Goble *et al.*, 2001) ; enfin l'approche d'intégration par lien (ou

dite de *mashup*) dans les projets SRS (Etzold & Argos, 1993) et Integr8 (Kersey *et al.*, 2005).

Toutes ces approches proposent des méthodes et des techniques pour résoudre des problématiques liées à l'accès à l'information en fonction de son lieu, mais pas nécessairement en fonction du contenu informationnel des données, à savoir de leur sémantique. La problématique de l'intégration de données grâce à la sémantique se pose dans un contexte plus général d'intégration qui est divisé en six couches dans Tolk (2006). Cependant, dans le cadre de l'intégration de données, nous nous limiterons aux trois premières couches d'interopérabilité : technique (réseau, couche d'accès logique aux données, APIs), syntaxique (type de données, terminologie) et sémantique (sens). Dans le domaine de la santé, l'interopérabilité sémantique de données biomédicales a été expérimentée dans le projet caBIG au travers de leur méthodologie semCDI (Shironoshita *et al.*, 2008). Leur approche n'a pas encore pu être validée correctement à cause d'un manque de formalisation de la connaissance de leur domaine (ontologies de domaine).

Notre travail s'effectue dans le contexte d'intégration de données biomédicales provenant d'un réseau d'hôpitaux européens dans le cadre du projet DebugIT[1] (Lovis *et al.,* 2008) (Detecting and Eliminating Bacteria Using Information Technology). Ce projet vise à intégrer des données cliniques et opérationnelles directement depuis les dossiers patients afin de proposer une vue globale de celles-ci à des fins d'analyse (datamining) et d'aide à la décision dans le domaine de l'antibiorésistance. L'accès à ces données distribuées et hétérogènes doit, pour des raisons de confidentialité des données, se faire de manière virtuelle et non matérialisée. Les données doivent donc être intégrées en temps réel.

Dans le domaine de la santé, l'utilisation grandissante de terminologies ou bien d'ontologies dans les systèmes de dossier patient ou de bases de données de recherche, nous a motivé pour valider l'utilisation des méthodologies et des technologies issues de la communauté du web sémantique. En particulier, la difficulté de représentation des données et des vocabulaires biomédicaux, pourrait mettre en exergue des limites dans l'utilisation des outils du web sémantique. C'est pourquoi nous proposons une méthode d'intégration pour le web sémantique en 3 couches (technique, syntaxique, sémantique) que nous expérimentons grâce à des outils sur des données biomédicales issues des systèmes opérationnels d'hôpitaux européens.

Dans la section suivante, nous présentons la méthode d'intégration proposée et validée dans le développement de la plateforme d'intégration de DebugIT. Nous présentons les résultats obtenus en section 3 et enfin, nous concluons en section 4.

2 Des données vers le web sémantique

Dans le cadre du projet DebugIT, diverses contraintes sont définies :

- Une vue unique et homogène des données du projet DebugIT doit être mise en œuvre.

[1] http://www.debugit.eu

- L'accès aux données doit être transparent, tant au niveau technique, syntaxique que sémantique.
- Les contraintes de confidentialité associées aux données ne nous permettent pas de stocker les données dans un entrepôt de données de manière centralisée. Chaque pays a d'ailleurs ses propres législations concernant la politique de confidentialité des données, et la méthode proposée devra pouvoir en tenir compte.
- L'accès direct aux données des dossiers patients depuis l'extérieur est donc généralement impossible pour des raisons de sécurité.

Ce sont les raisons pour lesquelles nous proposons de mettre en œuvre dans chacun des hôpitaux partenaires un entrepôt de données cliniques sémantique. C'est cet entrepôt (CDR[2]) qui sera visible de l'extérieur grâce aux technologies du web sémantique.

2.1 Sources de données

La première étape du projet vise à intégrer 4 sites. Les données partagées concernent le jeu de données « épisodes de soins ». Un catalogue de données a d'abord été crée en suivant une approche "bottom-up". Celui-ci contient les éléments d'information (artéfacts) requis pour répondre aux questions définies dans les cas d'utilisation du projet. Les catalogues de données des différents sites sont alignés. Les termes du catalogue partagé sont normalisés à l'aide de terminologies de référence comme SNOMED CT mais aussi NEWT et WHO-ATC. Chaque concept est donc enrichi d'une définition issue d'une ressource externe comme le MeSH ou wikipedia. Le jeu de données final représentant les « épisodes de soins » est agrégé dans des classes et des propriétés, comme pour la classe *Antibiogram* et ses propriétés *identified pathogen*, *antibiotic text*, *sensibility* et *identified pathogen concentration*.

Les questions d'experts posées aident à l'enrichissement de ce catalogue. Par exemple : "What is the **sensibility** of **pathogen** x found in **sample** y against **antibiotic** z over a **period** t?" ou bien "How to treat **disease** x caused by **pathogen** y for a **patient** with **z**?". Ainsi donc, les classes et propriétés peuvent être étendues à *Culture* (sample type), *Antibiogram* (identified pathogen, antibiotic tested et sensibility), *Patient treatment* (main/secondary diagnosis et commorbidity) et à la propriété *date.*

2.2 Méthode d'interopérabilité

Afin d'assurer une interopérabilité sémantique au sein de la plateforme DebugIT, nous proposons une méthodologie de mise en œuvre des CDR qui se décompose en 3 étapes (figure 1):

1. Technique – Concerne l'accès aux différents types de stockage de données. Les données opérationnelles peuvent être en texte libre, en format XML, dans des bases de données. Les modèles de données varient

[2] Clinical Data Repository

d'une source à l'autre, depuis du texte libre jusqu'à des systèmes normalisés grâce à HL7 ou OpenEHR.

2. Syntaxique – Les données sources sont rarement contraintes à un vocabulaire standard. Les termes sont souvent codés en texte libre, de manière abrégée ou avec des erreurs. Il est courant de trouver plusieurs occurrences du même terme noté de manière différente. Par exemple, la prescription d'antibiotique suivante : « TMP 300 mg 2/jour ».
3. Même si les données sont accessibles et partagent une partie de leur vocabulaire, seule une représentation formelle de chaque source de données et l'annotation de ces sources à une représentation formelle du domaine permet une interopérabilité sémantique des données. Par exemple, les données représentent des cas où la bactérie *E. Coli* est résistante à la *Trimétoprime*, la formalisation sémantique des données permettrait de répondre à la requête « Quels patients ayant *E. Coli* résistante à la *Péniciline ?*».

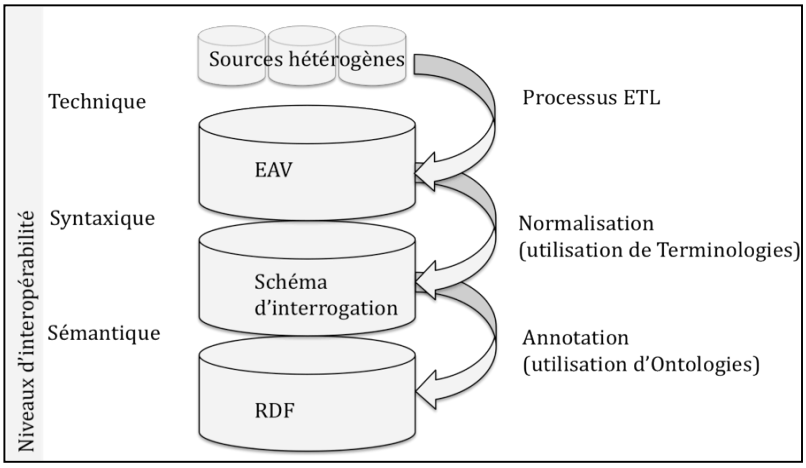

Fig. 1 – Les 3 couches de l'approche d'interopérabilité pour l'intégration de données de DebugIT

2.2.1 Interopérabilité Technique

Comme décrit dans la table 1, les sources de données présentent différentes plateformes techniques et différents protocoles d'accès aux données.

Nous proposons la mise en œuvre d'une couche technique et logique d'accès aux données à travers la mise en œuvre d'un CDR entre le système d'information hospitalier (SIH) et le réseau DebugIT. Cette couche intermédiaire de stockage de données est constituée par un espace de stockage permanent grâce à un SGBD. Le modèle de stockage de chaque CDR est de la forme EAV[3] (Nadkami *et al.*, 1999).

[3] Entité Attribut Valeur

Cette modélisation physique verticale permet de diminuer le coût de traitement au niveau de la base de données lors d'opérations telles que l'ajout de nouveaux concepts ou propriétés au catalogue de données commun. En effet, la représentation des propriétés en lignes (tuples) plutôt qu'en colonnes, permet des insertions/modifications de concepts sans modification du modèle physique de la base de données.

Source de données	Type de stockage	Système d'exploitation	Langue	#épisodes de soins
HUG	SGBD/texte libre	Windows	Français	20357
INSERM	SGBD	Windows	Français	3629551
LiU	Fichier csv	Linux	Suédois	103140
UKLFR	Texte Libre	Linux	Allemand	7949

Table 1. Description des bases de données sources (Hôpitaux Universitaires de Genève, Institut National de la Santé et de la Recherche Médicale, Linköpings Universitet, Universitätsklinikum Freiburg) : vue technique

Un processus ETL[4] est mis en œuvre entre le SIH et le CDR. Des agents d'extraction exécutent les tâches de chargement de données depuis le SIH, puis par des processus de transformation de modèle, chargent dans le CDR local. A cette étape, les sources de données DebugIT sont techniquement normalisées. La suite Talend[5] OpenStudio est utilisée pour le développement des agents d'extraction de données. Elle permet une représentation semi-automatique de la source de données (SIH) ainsi que, grâce à des modules inclus, de s'affranchir des problématiques d'accès à diverses sources de données (SGBD, XML, csv, etc.). De plus, elle permet une mise en œuvre plus aisée de la transformation de modèle SIH-CDR grâce à une interface utilisateur. Les CDR locaux sont alors mis en œuvre derrière la zone démilitarisée du SI de l'hôpital (DMZ).

2.2.2 Interopérabilité Syntaxique

Le contenu des sources de données (table 1) est multilingue (Français, Suédois et Allemand). De plus, des erreurs de codages sont fréquentes (ex : *Staph. Aureus* et *Staphylocoque aureus*). Chaque site ne présente pas le même niveau de normalisation par des vocabulaires contrôlés, c'est pourquoi nous proposons à cette étape un processus de normalisation, que ce soit pour résoudre la problématique du multilinguisme, ou bien, la problématique de la qualité de données.

Le processus de transformation des attributs codés dans un vocabulaire non contrôlé vers un vocabulaire standard et partagé est opéré par des agents de normalisation développés au niveau de l'EAV grâce à des routines écrites en PERL. Le processus est le suivant : 1) vérifier l'existence de données non normalisées dans le CDR ; 2) normaliser les objets trouvés ; 3) si succès, stocker la valeur trouvée ; 4) sinon, marquer comme échec. Certains objets comme la date sont convertis vers leur

[4] Extract Transform Load – Extraction Transformation Chargement

[5] www.talend.com

terminologie respective par les agents. Pour d'autres, comme les pathogènes ou les antibiotiques, des services tiers sont utilisés comme sources de normalisation. Dans le cas ou un concept peut prendre peu de valeurs différentes, par exemple le *sexe*, alors les valeurs sont liées manuellement au concept terminologique référent. Pour d'autres, comme par exemple les *bactéries*, un algorithme de fouille de texte simple a été mis en œuvre afin d'annoter et de normaliser les termes. Concernant les *antibiotiques*, l'algorithme tente d'abord de comparer les chaines de caractère avec WHO-ATC, si aucune correspondance n'est trouvé, alors une lettre sera substituée, et ainsi de suite.

2.2.3 Interopérabilité Sémantique

Trois étapes sont définies afin de réduire la distance entre les données opérationnelles et les représentations formelles du domaine (ontologies de domaine). Premièrement, la base de données source est définie formellement (Data Description Onotlogy : ontologie de données) suivant deux axes : le modèle de données et le vocabulaire. Deuxièmement, une représentation partagée des concepts du domaine est créée (Schober *et al.*, 2010) (DebugIT Core Ontology : ontologie de domaine). Enfin, un lien entre la représentation formelle de la base source et les concepts du domaine est mis en œuvre à travers un médiateur de requêtes basé sur des règles.

Un nombre important d'approches pour exposer des données en RDF sur le web de données (ou *LinkedData*) ont été proposées dans la littérature (Broekstra *et al.,* 2001 ; Openlink[6]). Cependant, ces approches ne permettent pas de faire de l'intégration de données (Bizer & Cyganiak, 2007). Nous ne détaillerons pas dans cet article la composante intégration de données au niveau sémantique qui est représenté par la troisième étape précédemment citée, c'est un sujet à part entière qui fera l'objet d'un article ultérieur.

D2RQ[7] est la couche *middleware* choisie dans le cadre de DebugIT afin de transformer les données relationnelles en RDF. La génération automatique de *mapping* proposée par D2R transforme chaque table en classe et chaque colonne en propriété. Ce fichier de *mapping* est utilisé par Jena afin de transformer les requêtes SPARQL en requêtes SQL. L'approche est simple mais limitée puisqu'il n'est pas possible de gérer les problématiques de vocabulaires pour l'intégration. Nous proposons donc d'intégrer une formalisation de la base de données dans un formalisme ontologique, la DDO. Cette ontologie a pour but de décrire non seulement les classes et propriétés de la base de données, mais aussi le vocabulaire dans lequel les instances d'une propriété sont stockées.

Enfin, la DCO, qui contient à ce jour 894 classes incluant 294 classes BioTop (Schulz *et al.,* 2006), sera utilisée pour exprimer les requêtes faites sur les différents CDR/D2R. Notre approche de médiation de requêtes se basant sur l'opérateur CONSTRUCT de SPARQL qui permet de créer un graphe à partir d'autres graphes sources de l'opérateur WHERE.

---

[6] OpenLink Software, "Virtuoso: Universal Server Platform for the Real-Time Enterprise. http://www.openlinksw.com/virtuoso/

[7] The D2RQ Platform - Treating Non-RDF Databases as Virtual RDF Graphs, http://www4.wiwiss.fu-berlin.de/bizer/d2rq/

3 Résultats

Les quatre sites ont été intégrés suivant les trois couches d'interopérabilité définies dans notre méthodologie.

L'interface de *mapping* fournie par Talend ainsi que les routines de verticalisation du modèle de données (relationnel vers EAV) ont été efficientes pour la partie technique. La table 2 représente un extrait de la table EAV d'un CDR. L'EAV apporte une grande flexibilité vis à vis du modèle de données. En effet, si le modèle vient être étendu, il n'est pas nécessaire de modifier le schéma, et donc, d'arrêter le système. Par exemple, rajouter un triplet "culture#5579709, sample_type_location, urine" rajoute un concept *sample_type_location* et la valeur *urine* à la culture *#5579709* sans avoir à modifier la table (ALTER TABLE).

L'intégration syntaxique est effectuée afin de répondre aux questions d'experts posées en section 2.1. Les agents normalisateurs pour les domaines des antibiotiques et des bactéries peuvent être partagés entre les CDR. Cependant, certaines spécificités encouragent chaque partenaire à développer ses propres agents normalisateurs. La table 3 représente le pourcentage de normalisation de chaque site source en fonction des objets du domaine.

Entité	**Attribut**	**Valeur**
culture#5579709	Episode_of_care_id	10744189
culture#5579709	Collect_date	2007-12-14 11:20:00
culture#5579709	Result_date	2007-12-18 11:58:00
culture#5579709	Culture_procedure	Culture aérobie
culture#5579709	Antibiotic_tested	AM.CLA
culture#5579709	Antibiotic_tested_result	S
culture#5579709	Identified_bacteria_name	Escherichia coli
culture#5579709	Identified_bacteria_quantity	10^E5

Table 2. Extrait de données issues d'un modèle EAV pour la culture #5579709 de HUG

Propriété	**Terminologie**	**Statut de normalisation**			
		HUG	INSERM	LiU	UKLFR
Date	Time.OWL	100%	100%	100%	100%
Pathogen	NEWT	97%	86%	95%	100%
Antibiotic	WHO-ATC	98%	80%	100%	100%
Sample Type	SNOMED CT	100%	80%	100%	100%
Diagnosis	ICD-10	100%	100%	-	0%
Comorbidity	ICD-10	100%	100%	-	0%
Culture Procedure	SNOMED CT	99%	93%	-	100%
Sensibility	SNOMED CT	100%	100%	50%	100%
Treatment frequency	SNOMED CT	94%	100%	-	88%

Table 3. Statistiques du taux de réussite du processus de normalisation pour chaque site

La normalisation s'est montrée plus complexe à effectuer que l'extraction des données vers l'EAV. En effet, le manque de terminologies standard et consensuelles ainsi que la mauvaise qualité des données contenues dans des systèmes opérationnels qui n'implémentent pas de vocabulaire standardisé pour coder l'entrée des données par les praticiens hospitaliers, rendent difficile la normalisation des données à des fins de partage sur le web de données. La table 3 montre cependant que les résultats obtenus restent satisfaisants.

Bien que les technologies du web sémantique permettent, en partie, d'effectuer une intégration technique (via le protocole SPARQL) et syntaxique (*mappings* dans D2R), celles-ci restent limitées et non généralisables. C'est la raison pour laquelle nous proposons la mise en œuvre des CDR via les deux étapes précédentes. De plus, les deux étapes précédentes permettent de faciliter la mise en œuvre de la couche sémantique du CDR. En effet, la normalisation effectuée aide à l'annotation des données avec la DCO (ontologie de domaine). La première version des SPARQL *endpoint* mis en œuvre dans le cadre du projet DebugIT sont annotés manuellement à la DCO. La validation de l'approche d'intégration sémantique du démonstrateur sera effectuée avec l'exécution des requêtes SPARQL sur les quatre SPARQL *endpoint*. Le langage SPARQL permet de construire des graphes avec des concepts issus de la DCO, en fonction des concepts issus des DDO.

La méthodologie de médiation adoptée ici est appelée *Global as View* où l'ontologie DCO est appliquée comme une vue au dessus des données sources. Une application médiatrice de démonstration a été écrite afin d'exécuter la requête sur les 4 endpoints accessibles. Les résultats de cette requête ("*What is the **sensibility** of **E.Coli** found in **urine** against **Trimethoprim?")*** sont présentés dans la table 4. Ce système de médiation basé d'une part sur la réécriture des résultats via la clause CONSTRUCT et la réécriture de la clause WHERE via des règles seront développés dans un article ultérieur dans le cadre de la mise en œuvre de la plateforme d'interopérabilité (IP) (Choquet *et al.*, 2009).

SPARQL endpoint	**Temps de réponse**	**#triplets**
https://babar.unige.ch:8443/d2r-server/sparql	6535 ms	17072
http://debugit1.spim.jussieu.fr/sparql	6522 ms	34276
http://lincoln.imt.liu.se:2020/sparql	1008 ms	1518
https://codeine.medinf.uni-freiburg.de:8443/debugIT/sparql	495 ms	0

Table 4. Nombre d'enregistrements retournés pour l'exécution d'une requête SPARQL sur 4 *endpoints.*

Les résultats obtenus seront traités par des modules d'aide à la décision dans le cadre du projet DebugIT. Sur la question concernant la sensibilité de la Trimethoprim à la bactérie E.Coli trouvé dans des cas d'infection urinaires, les dataminers seront capable, grâce aux résultats formalisés avec la DCO, de générer de nouvelles connaissances non ambigües et pourront utiliser les mécanismes d'inférence et de raisonnement propres aux graphes RDF.

4 Discussion et conclusion

La demande grandissante de partage de données biomédicales, en temps réel, à des fins d'amélioration des soins ou de la prise en charge du patient nous pousse à nous interroger sur la viabilité des technologies de l'information proposées par le W3C et plus particulièrement le groupe de travail sur le web sémantique pour le traitement de l'information biomédicale. Le projet DebugIT vise, entre autres, à valider l'hypothèse que ces outils sont viables méthodologiquement et techniquement (par ex : montée en charge). Nous avons expérimenté cette méthodologie dans le cadre d'un projet européen DebugIT sur 4 centres hospitaliers fournisseurs de données hétérogènes tant au niveau technique, syntaxique que sémantique.

Nous avons proposé une méthodologie d'intégration de données opérationnelles biomédicales basée sur les trois premières couches de l'interopérabilité des systèmes d'information. Nous pensons cette méthodologie nécessaire dans le cadre du partage de données issues de systèmes d'informations hospitaliers sur le web sémantique, surtout dans le cas où les données à partager sont issues des systèmes opérationnels tels que le dossier patient. Notre méthode d'intégration permet de scinder les problématiques en différentes étapes. En effet, il est difficile de croire aujourd'hui qu'un problème d'intégration de données opérationnelles peut être résolu simplement à la dernière étape de notre processus (intégration sémantique). Ceci pour des raisons de mauvaise qualité des données, d'hétérogénéité de représentation de l'information ainsi que pour la confidentialité des données médicales. De plus, les outils du web sémantique pour partager des données sur le web ne permettent pas une réelle intégration (et intégration sémantique) des bases de données. En effet, il est difficile d'interroger des données RDF issues de bases de données à partir d'ontologies où chaque élément est défini comme un concept, alors qu'une base de données est composée essentiellement d'instances.

Le partage de données médicales sur le web pose aussi des problèmes réglementaires importants et différents suivant les pays. La problématique de l'anonymisation des données est abordée dans le cadre du projet DebugIT et doit l'être de manière plus globale dans le cadre du partage de données au niveau du web sémantique. Nous partageons aujourd'hui nos données de manière anonyme et de manière cryptée sur le web. Il conviendra d'intégrer dans toute plateforme de partage de données médicales des processus de sécurité fiables.

Nous devons compléter dans un travail en cours la partie sémantique de notre système intégration afin d'offrir à l'utilisateur final un accès transparent aux données de tous les centres participants au réseau DebugIT au niveau sémantique. Ce travail sera publié prochainement et présente déjà des résultats prometteurs.

References

BIZER C. and Cyganiak R. (2007). D2RQ Lessons Learned. In: W3C Workshop on RDF Access to Relational Databases, W3C Workshop on RDF Access to Relational Databases.

BROEKSTRA J. et al. (2001) "Sesame: An Architecture for Storing and Querying RDF Data and Schema Information", http://www.cs.vu.nl/~frankh/postscript/MIT01.pdf

CHOQUET R. et al. (2009). Specifications of an Inter-Operability Platform for the integration and exploitation of distributed clinical data. American Medical Informatics Association Annual Symposium Proceedings.

CHOQUET R et al. (2010). The Information Quality Triangle: a methodology to assess Clinical Information Quality. MEDINFO 2010.

ETZOLD T. and ARGOS P. (1993). SRS – an indexing and retrieval tool for flat file data libraries. COMPUTER APPLICATIONS IN THE BIOSCIENCES 9, 49-57.

GALPERIN MY. (2008) The Molecular Biology Database Collection: 2008 update. Nucleic Acids Research 36. D2–D4.

GOBLE C. A. et al. (2001). Transparent access to multiple bioinformatics information sources. IBM SYSTEMS JOURNAL 40, 532-551.

IAVINDRASANA J. et al. (2007). Design of a Decentralized Reusable Research Database Architecture to Support Data Acquisition in Large Research Projects. Stud Health Technol Inform 129, 325-9.

KARASAVVAS KA et al. (2004). Bioinformatics integration and agent technology. Journal of Biomedical Informatics 37, 205–219.

KERSEY P. et al. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. NUCLEIC ACIDS RESEARCH 33, D297-D302.

LEE T. J. et al. (2006). BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics 7, 170.

LOVIS C. et al. (2008). DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Tech Inform 136 (2008), 641-6

NADKARNI P.M. et al. (1999) Organization of heterogeneous scientific data using the EAV/CR representation, J Am Med Inform Assoc. Nov-Dec;6(6):478-93.

PILLAI, S. V. et al. (1987). Design issues and an architecture for a heterogenous multidatabase system. Proceedings of the 15th ACM Computer Science Conference.

SCHOBER D. et al. (2010), The DebugIT Core Ontology: semantic integration of antibiotics resistance patterns. MEDINFO 2010.

SCHULZ S. et al. (2006)., Towards an upper level ontology for molecular biology. American Medical Informatics Association Annual Symposium Proceedings.

SHETH AP and LARSON JA. (1990). Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. ACM Computing Surveys 22, 183-236.

SHIRONOSHITA et al. (2008). semCDI: a query formulation for semantic data integration in caBIG. Journal of the American Medical Informatics Association vol. 15 (4) pp. 559-568

TOLK A. (2006). What Comes After the Semantic Web - PADS Implications for the Dynamic Web, 20th Workshop on Principles of Advanced and Distributed Simulation (PADS'06)

TÖPEL T. et al. (2008). BioDWH: A Data Warehouse Kit for Life Science Data Integration. Journal of Integrative Bioinformatics 5, 93.