DOI: 10.15439/2015F302
ACSIS, Vol. 5

Multisource agent-based healthcare data gathering

Vincenza Carchiolo, Alessandro Longheu, Michele Malgeri and Giuseppe Mangioni

Dip. Ingegneria Elettrica, Elettronica e Informatica - Università degli Studi di Catania - Italy
{vincenza.carchiolo, alessandro.longheu, michele.malgeri, giuseppe.mangioni}@dieei.unict.it

***Abstract*—The number and type of digital sources storing healthcare data is increasing more and more, rising the problem of collecting actually dispersed information about a single patient. In this paper we propose an agent-based system to support integration of health-related data extracted from both structured (HIS) and semi-structured (websites and social networks) sources. Integrated data are exported in HL7 format to finally feed personal health record (PHR).**

I. Introduction

DIGITAL healthcare data dramatically increased during last years [1]; this is mainly due on one hand to data stored into Health Information Systems (HIS) [2][3] as medical records and clinical exams, while on the other hand a significant contribution comes from specialized websites (e.g. online medical forums) and social networks as FaceBook and Twitter, where doctors and patients post their personal experiences and opinions about various health related topics e.g. illnesses, symptoms, treatments, side effects [4][5].

The number and type of such data sources poses the problem of collecting actually dispersed healthcare information for a single patient, therefore the extraction and integration of such data into a standard repository arise.

Data integration is a well-known and quite old issue [6][7], however it still requires a significant effort, especially when information are extremely sensitive for privacy issues, as it occurs for medical information [8].

In this paper we propose an agent-based systems whose goal is the integration of health related data coming from different sources; as cited before, we consider both HIS (i.e. SQL based data sources) as well as websites (HTML-based data sources) and social networks (in particular, Twitter).

These are used to populate official medical documentation known as Electronic Medical Records (EMR), Electronic Health Records (EHR) and Personal Health Record (PHR). In more detail, EMR/EHR [9][10][11] is the digital collection of a person's health related documents used within HIS to provide an effective, reliable and costs saving health management, contains the standard medical and clinical data and is managed only by health care providers, whereas in the PHR [12] the person can directly manage his personal medical-related information and therefore it can also contain data coming from website and social networks cited previously (i.e. unstructured or semi-structured). In the rest of this paper, we will use the term PHR only, implicitly including also EMR/EHR.

Since our goal is the gathering of *personal* health-related data, we suppose that all information can be accessed on a per-user basis according to authentication mechanisms whenever present, i.e. agents should be granted access to personal profiles to extract data.

Moreover, for an effective integration a common reference terminology is needed; to this purpose, we exploit the SNOMED-CT [13], currently the most comprehensive medical terminology worldwide adopted, to discover medical terms and properly manage, match and integrate synonyms, hyponyms and hypernyms.

Together with standard database terminology, we also include a list of additional informal terms to be searched if nothing is found within SNOMED-CT, indeed it is possible that within unofficial medical data sources (as in the case of social networks and websites) people use words like ”headache” that are not explicitly stored into SNOMED-CT, where ”headache” is indeed referred as ”migraine”; the additional list allows to cope with such real situations.

Finally, a common output format for PHR data is advisable; a widely accepted format is HL7 [14], a standard for information exchange between medical applications and healthcare providers. HL7 includes several recommendations for conceptual representation, documents (included the PHR), applications and messaging standards, and is available as v2.x and v3.0. The v2.x version is a non-XML proposal where data is organized in segments (lines), each containing proper fields and subfields. The v3.0 HL7 messaging protocol leverages XML to provide data structure and also provides support for healthcare workflows.

Other works deal with integration issues, for instance in [15] an OWL ontology is developed to integrate specific medical documents (CCD) authors focus on, whereas in [16] a complete method for ontology based schema and data integration for clinical and genomic databases is presented. Our goal is to provide a tool for extracting and integrating any (neither specific nor structured) medical information without creating a new ontology (rather, exploiting existing ones). Using social forums in healthcare has been investigated e.g. in [17], where answer for medical queries in unresolved posts is provided via similar thread retrieval; to the best of our knowledge, no data gathering from social forum for PHR has been considered so far. Joining virtual social networks and healthcare recently led to neologisms as *Infodemiology* and *Infoveillance* [18]; also Twitter has been exploited in this sense, for instance in [19] the micro-blog is used to detect flu trends, whereas in [5] Ailment Topic Aspect Model is applied to tweets to track public health over time; in [20] authors tracked and

978-83-60810-66-8/$25.00©2015, IEEE

examined disease transmission in particular social contexts via Twitter data, while in [4] social media use improves healthcare delivery by encouraging patient engagement and communication. Apart all these proposals however, no specific use of Twitter data for PHR currently exist. A preliminary and partial study of the work presented in this paper can be found in [21] and [22].

The paper is organized as follows. In section II we describe the overall architecture of our proposal, and how the data extraction and integration are performed for each data source. In section III we show an application to a real case, providing concluding remarks and future works in section IV.

II. Agent-based gathering system

In fig. 1 our agent-based model is shown. The three main data sources categories we consider are the so-called *database*, that represents standard HIS where official medical personal data are stored (usually, according to a well structured schema), and the *website* and *social*, indicating respectively HTML-based and text-based data sources as described in previous section; in our experiments in particular we looked at online medical forums for the ”website” category and Twitter for the ”social”. Considering agents, the first set (named *wrapper agents*) is devoted to gather personal data from each data source, while *text mining agents* filter previous data to extract medical information. The integration module collect all such data and performs proper integration also exploiting both the SNOMED-CT terminology and an additional dictionary of ”common” medical terms (e.g. ”headache”) that are not stored as official medical terms in the SNOMED-CT. The same references are used by the text mining agents to detect medical information. After this, *feeding agents* are used to populate user's PHR with his relevant medical data. *User agents* collaborate to compare information gathered from different users but semantically related, allowing to build a user network; similarly, *disease agents* cooperate to correlate detected diseases. In the following each component of the proposed architecture is described in more detail.

A. Wrapper agents

All sources are managed by wrapper agents, that are used to isolate and collect information referring to the same user; in the case of database this is quite trivial and can be accomplished via standard SQL based queries, even if each real HIS has its schema and probably both proprietary solutions as well as authentication issues must be tackled. The question is somehow different for what concern websites and social network. In the case of websites indeed, in particular considering online medical forums we focused on, sometimes all messages are directly available on a per-user basis, therefore the extraction performed by the agent still remains feasible, whereas when forums provide a thread/topic based classification the wrapper agent has to browse all threads/categories and collect all messages for each single user to contribute as much as possible in populating his medical profile; forums where total anonymous messages are allowed are then not considered here (i.e. registered nicknames only). In social networks as Facebook or Twitter, the same approach can be applied since they generally adopt the same organization of traditional websites, i.e. personal messages somehow identifiable even in the case of a topic-based arrangement. Note that the term *person* here refers to a distinguishable user_id or nickname, i.e. the agent stores data together with the *(user_id, source_id)* pair so that each text refers to a single user. Since the nickname should be associated to a real person, this requires he/she should grant the permission for his/her data in order to fulfill privacy issue and related laws; here we suppose that persons do not prevent their personal data to be extracted by agents.

B. Text mining agents

The next step is the extraction of all medical related data (concerning a given person), discarding other information; this task is performed by text mining agents shown in fig. 1. This agent is not present in the case of HIS since it is likely that such a data source exclusively stores medical related information; conversely, websites and social networks may also include non medical data, also thanks to their semi-structured or unstructured nature, i.e. HTML and text based data posted by common people (not necessarily doctors or medical personnel). Online medical forum and social networks are managed following the same approach, except for a preliminary step in websites, where pure text is extracted from HTML, therefore these sources both provide text data (either webpages or tweets). The first operation we apply is the language recognition, since all further text-mining actions (e.g. stemming) strictly depend on the specific language; in this sense, we discard then all non English text portion. Then, the text mining agent searches for statements containing medical terms; this is accomplished using Natural Language Processing (NLP) techniques [23], in particular first removing irrelevant information (as hyperlink text for web pages, or retweet details and usernames for tweets) and then applying standard text processing operations as tokenization, stopwords removal, stemming and indexing [24]. If the index terms list contains at least a medical term, the statement that term belongs to is preserved, otherwise we discard the statement.

In the last step, we leverage sentiment analysis [25][26] in order to evaluate the remaining statements (those containing medical terms), e.g. if the person suffers a disease cited in a tweet, or if an intolerance to some food specified in a forum question is still present or not. Sentiment analysis or opinion mining [27] leverages NLP, text analysis and computational linguistics to extract subjective information, as the mood of the people regarding a particular product or topic; basically, the sentiment analysis can be viewed as a classification problem of labelling a given text (statement within a tweet or extracted from online forum) as *positive*, *negative* or *neutral*, in our case if the result of classification is positive or negative, the text mining agent passes this information to the integration module to enrich personal medical profile.

To understand how it works, let us consider the text ”Last night was too rainy, this morning my headache is stabbing but

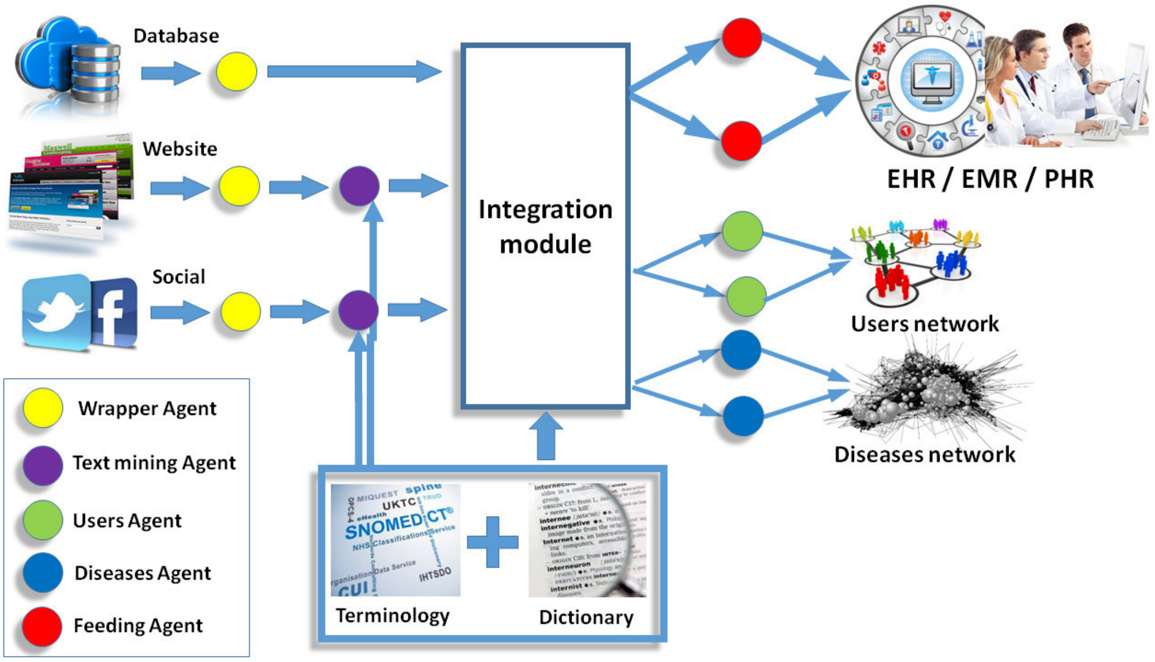


Fig. 1. Application architecture

fortunately I will not go at work" that can come from a social as well as from a website source. Using the NLTK chunking package [28], we first identify short phrases (clusters) like noun phrases (NP) and verb phrases (VP); the text portion cited above in particular produces the following chunks:

"Last night"(NP)
"was" (VP)
"too rainy" (NP)
"this morning" (NP)
"my headache" (NP)
"is stabbing" (VP)
"but fortunately" (NP)
"I" (NP)
"will not go" (VP)
"at work" (NP).

The sentiment analysis exploits the chunking technique first to isolate clusters, then we discard those without medical terms (in the example only "headache" is present, therefore clusters "Last night was too rainy" and "I will not go at work" are discarded), finally trying to assess the meaning of remaining clusters by using a proper list of *positive* and *negative* verbs, so the cluster "this morning my headache is stabbing" is labelled as *positive* or, in other words, we got the information that person has the headache.

C. SNOMED-CT and dictionary

In order to discard statements that do not contain any medical term, we search each index term extracted by text mining agent in the SNOMED-CT terminology. To better clarify how this search is performed, we briefly cite the SNOMED-CT core components (details can be found in [29]) that are:

- *concepts*, that represent all entities that characterize health care processes; they are arranged into acyclic taxonomic hierarchies (according to a *is-a* semantics)
- *descriptions*, explaining concepts in terms of various clinical terms or phrases; these can be of three types, Fully Specified Names (FSNs) that is the main (formal) definition, Preferred Terms (PTs), i.e. the most common way of expressing the meaning of the concept, and Synonyms.
- *relationships* between concepts, e.g. the concept (disease) "Staphylococcal eye infection" has "Causative agent" relationship with "Staphylococcus" (different types of relationships exist depending on concepts type)
- *reference sets* used to group concepts e.g. for cross-maps to other standards purposes.

In this work, the first two items are considered, in particular among all concepts hierarchies we focus in the "disorder/disease" since our goal is to detect tweets about diseases, therefore we do not consider other specific hierarchies (e.g. "surgical procedures"). Inside the disorder hierarchy, we search each index term extracted from tweets as a FSN, PT or synonym; if found, that tweet is further processed in order to establish whether the specified disorder is present using sentiment analysis (see below).

As indicated in the introduction, to guarantee that all medical terms can be successfully detected, a list of additional informal terms is searched if nothing is found within SNOMED-CT, for instance if the index term is the word "flu", this has positive match in the synonym list of "influenza" disease (the FSN), but the (also quite common) term "headache" is not explicitly present when browsing SNOMED-CT [30], where this disorder is instead referred as ”migraine” both as FSN and its synonym. Including ”headache” in an additional list (*dictionary* in fig. 1) is the simple solution we adopted; this list is considered just if nothing is found within SNOMED-CT.

Also note that several diseases are defined as a group of words (e.g. ”Viral respiratory infection”), therefore during the indexing phase we also retain N-grams with N=2 and 3; diseases with more than three words can be easily disambiguated even with 3 words since not all words are generally significant (e.g. in ”Disease due to Orthomyxoviridae” the first and the last words are enough for correct matching).

Finally, detected diseases may be hierarchically related, e.g. ”influenza” and ”pneumonia” are both children of ”Viral respiratory infection” according to the ”is-a” semantics. This information could be used for instance by replacing both children with their common parent, in order to build a more generalized, global view of diseases named in the given geographic area during the chosen time period. We choose however to preserve the best level of detail by not using a common ancestor as in the example, while on the other hand we will substitute all terms that represent the same disease with its FSN as indicated in SNOMED-CT, for instance if different tweets refer to ”flu”, ”grippe” and ”influenza” they will be all considered as tweets about ”influenza”.

D. Integration module

The integration of all data concerning the same person into his PHR is performed by the *integration module* after it receives data by wrapper and/or text mining agents. The process is not fully automatic at this stage, so the scenario the user is presented to includes a set of pieces of information that could refer to the same person and have to be somehow integrated.

In particular, the wrapper (or text mining for semi-structured data) agents deliver a set of triples *(user_id, source_id, data)* each referring to a given user, though distinct user_ids, say *UID1* and *UID2* could refer to the same physical person since such triples come from different sources. Whenever *UID1* is literally identical to *UID2*, the system suggests to integrate all related *data* into the same set for further processing, and the user simply can confirm this decision; if however the user believes that although identical those pieces of information do not refer to the same person, for instance when the first *data* is about menopause (female) and the second is about prostate cancer (male), the system stores this decision and rename the second id differently (i.e. *UID2_*) in order to allow disambiguation for further data. In the case *UID1* and *UID2* are different, the user has to decide whether these ids actually represent the same person or not; again, the system stores the decision, therefore if *UID1* is ”Robert Stanton” and *UID2* is ”RBTSTN” and the user establishes that the latter is a portion of ”Robert Stanton”’s fiscal code, all further triples having ”RBTSTN” will automatically incorporated into the same set of ”Robert Stanton” pieces.

After all pieces of health-related information belonging to the same person have been collected into the same set *S*, the integration module tries to integrate *data* present in triples, again through a semi-automatic process. It is important to highlight however that a definitive standard for PHR currently does not exist at all, therefore we do not aim at defining a schema all data should adhere to, rather our goal is to integrate to some extent data concerning the same information.

In particular we first focus on structured data, where the presence of a schema means that *data* in a triple is usually represented as a table, i.e. a query result coming from HIS (databases). Different tables can be integrated first from a structural point of view (e.g. merging ”patient_ID” and ”patient_CODE” columns into a single one) and then from a semantic perspective (e.g. collapsing ”migraine” and ”headache” into the same term); the problem is quite complex but really not new and several solutions already exist [31] [6] [7]. In our first implementation, the user can specify whether he wants a single schema or not; if so, for each table the user has to select which columns are considered (or discarded) and whether columns from different tables must be joined together into a single one; the data type of an output column that mixes two or more existing columns will be the largest data type among those of existing columns being integrated (e.g. float and integer are integrated to float). At the end of the process, a single table comes as output of the integration module and its definition will be XML based, according to HL7 v3.0 format cited in the introduction. Note that data belonging to tables are not integrated in our implementation, rather we simply insert all pieces of information into the single table; if the user do not wish a real integration, all tables are simply preserved and passed as output.

Considering the case of semi-structured data, i.e. website and/or social in fig. 1, *data* stored in a triple is not a relational-like tables, rather it is a labeled statement where the person (identified by the *user_id* element) suffers a given disease (as the headache in the example cited in the text mining module) or has been screened with a given clinical examination etc. In this case the integration we implemented allows to collapse several pieces of information whenever they actually coincide, for instance if we have the statements ”this morning my headache is stabbing” and ”the diagnosis was acute migraine” both labelled as positive and belonging to the same set *S* concerning a given person *user_id*, the system allows the user to discard one of the two statements since they represent the same information. Note that the system recognizes the statements as comparable since it exploits SNOMED-CT and the dictionary of informal terms to map all medical terms (in this case, ”headache” and ”migraine”) into the same term (in this case, the FSN ”migraine”); if such a mapping does not occur, statements are supposed to be different, but the system always allows the user to collapse a set of statements into a single one manually if this is the case.

E. Feeding agent

After the integration process has been performed, a (possibly reduced) set of triples *(user_id, source_id, data)* is provided by the integration module for PHR feeding. The way

PHR is actually stored and accessed has not been definitively standardized [32], however in our architecture PHR should be available as web services to guarantee an easy and uniform access. In this sense, a number of platforms have been proposed, as the Microsoft's HealthVault [33], PatiensLikeMe [34] or other proprietary solutions, in addition to institutional approaches, e.g. [35]. Since each one of them has its features, we specify only general guidelines of the actual implementation of the feeding agent, in particular each agent is devoted to a specific solution and has to manage authentication issues as well as data format and communication protocol; we suppose however that supporting HL7 v3.0 as XML-based data format is the best choice

F. User and disease agents

Users agents are devoted to build and manage the network of users whose data have been extracted. The network can be built according to different criteria, for instance two users may be considered as *linked* if they "share" the same disease, or they were admitted at the same hospital. Similarly, a set of *Diseases agents* can build a network of diseases somehow related, e.g. a link between two diseases may represent a co-occurrence of both diseases in a significant number of data concerning the same user, or the fact that they are cited by related users and so on; the establishment of such network leverages the SNOMED-CT to tackle semantic-related issue (e.g. synonyms, homonyms, hypernyms). At this stage these two set of agents are considered for future works and have not been implemented.

III. Results

In this section we show an example of how the architecture works. In particular, we considered a small group of persons who suffer different diseases and are in treatment at the same medical center; in fig. 2 we show one of them with two so-called *medical problems*, i.e. diabetes and osteoporosis. The Health Information System used in this centre was actually a customized version on the OpenEMR software [36], whose data access were granted to the wrapper agent through an API-based connection to the underlying MySQL database. The wrapper agent submitted an SQL query with users names, in order to get information about their exams; in fig. 3 we show the resulting table for the same person of fig. 2; for the sake of simplicity, we omitted query details, i.e. the foreign keys used to join patient with medical problems tables, and only significant columns are shown in fig. 2 (patient name and related medical problems). As specified in the previous section, the table will be the *data* field in the triple *('Rebecca Greenfield', 'SRC#002', data)* extracted from MySQL database for that person by the wrapper module and delivered to the integration module. Note that to get results, we supposed (as specified in the introduction) that the person in fig. 2 grants the wrapper agent to access her data, in order to overcome authentication issues.

Similarly, other wrapper agents search on websites and/or social networks, in particular we considered [37] (a medical forum) as "website" data source, and, supposed that we are able to associate (even manually) the name stored in the HIS to the nickname "rebgreen46", we allow the wrapper agent to extract from HTML the text contained within posts (fig. 4 shows one of them). The text is further processed by the text mining agent, as discussed in previous section, therefore we get the following triples:

- *('rebgreen46', 'SRC#003', ('back', 'hurts', POSITIVE))*
- *('rebgreen46', 'SRC#003', ('legs and shoulders', 'aches', POSITIVE))*
- *('rebgreen46', 'SRC#003', ('pain', 'is getting worse', POSITIVE))*

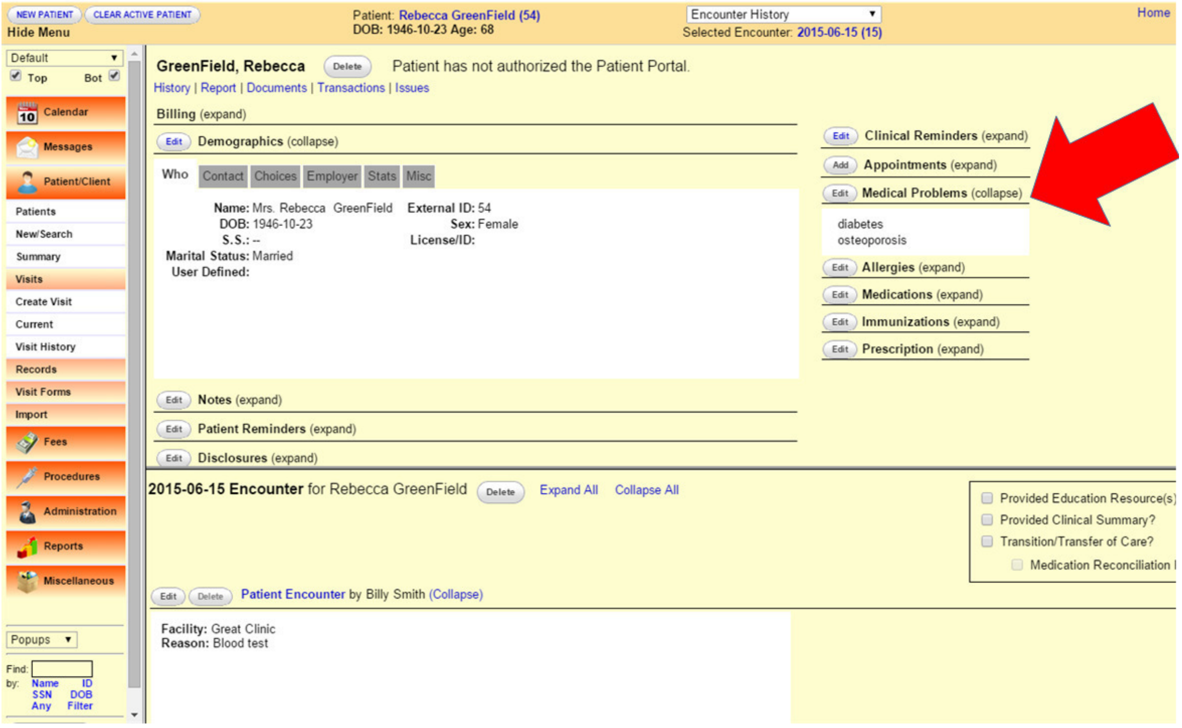
These triples represent therefore additional pieces of information about that person, i.e. from the database emerges that she suffers diabetes and osteoporosis, whereas from this forum the presence of aches for back, legs and shoulders is detected. According to the procedure described in the previous section, the integration module allow to associate these three triples with the table (in this case, manually specifying that 'Rebecca Greenfield' is the same *user_id* as 'rebgreen46'); one or more feeding agents will finally connect to the platform(s) where these pieces of information will be added to the Personal Health Record for that patient. Apart the simple example shown so far, we are currently undergoing on a more significant test with a relevant number of patients (about 200) also considering Twitter as "social" data source.

IV. Conclusions

The work described in this paper outlines an agent-based architecture for feeding PHR with data extracted from structured and semi-structured (databases and website/social networks, respectively) sources. Our proposal is currently at an early stage, and we believe that the first implementation will provide us with results confirming prosiming expectations. Moreover, the implementation will allow to assess the effectiveness of our approach, expecially when a relevant number of data sources as well as a relevant number of patients to extract data about will be gathered; this validation is also required to evaluate the quality of integrated infomation with respect to that stored into original sources (e.g. redundancy, completeness and so on).

Some future works are planned:

- a useful extension concerning the integration module is the possibility of specifying a mapping function so that the system can be trained a-priori to associate different *user_id* into the same person, for instance the mapping function can be a regular expression acting as a bijective function to convert *UID1* to *UID2* and viceversa. More complex function could be defined to allow automatic integration.
- the integration module presented in this paper is quite elementary; we are considering more effective solutions, e.g. a classifier (supervised or not) that tries to aggregate health data without the user's intervention or reducing this as much as possible. Besides, more effective data


NEW PATIENT
CLEAR ACTIVE PATIENT
Hide Menu
Encounter History
Home
Default
Top
Bot
Calendar
Messages
Patient/Client
Patients
New/Search
Summary
Visits
Create Visit
Current
Visit History
Records
Visit Forms
Import
Fees
Procedures
Administration
Reports
Miscellaneous
Popups
Find:
by: Name ID SSN DOB Any Filter
Delete
History | Report | Documents | Transactions | Issues
Billing (expand)
Edit Demographics (collapse)
Who Contact Choices Employer Stats Misc
Edit Clinical Reminders (expand)
Add Appointments (expand)
Edit Medical Problems (collapse)
Edit Allergies (expand)
Edit Medications (expand)
Edit Immunizations (expand)
Edit Prescription (expand)
Edit Notes (expand)
Edit Patient Reminders (expand)
Edit Disclosures (expand)
Delete
Expand All
Collapse All
Edit Delete
(Collapse)

Patient: Rebecca GreenField (54)
DOB: 1946-10-23 Age: 68

Selected Encounter: 2015-06-15 (15)

GreenField, Rebecca

Patient has not authorized the Patient Portal.

Name: Mrs. Rebecca GreenField
DOB: 1946-10-23
S.S.: --
Marital Status: Married
User Defined:

External ID: 54
Sex: Female
License/ID:

diabetes
osteoporosis

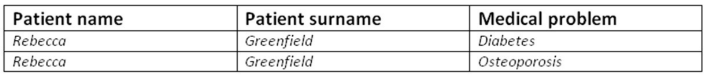
2015-06-15 Encounter for Rebecca GreenField

☐ Provided Education Resource(s)
☐ Provided Clinical Summary?
☐ Transition/Transfer of Care?
☐ Medication Reconciliation I

Patient Encounter by Billy Smith

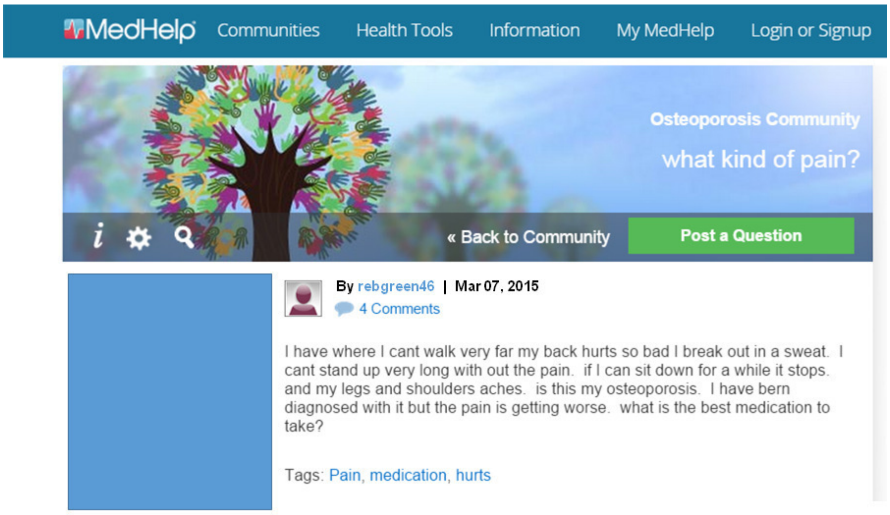
Facility: Great Clinic
Reason: Blood test

Fig. 2. Snapshot from HIS (structured data source)



Patient name	Patient surname	Medical problem
Rebecca	*Greenfield*	*Diabetes*
Rebecca	*Greenfield*	*Osteoporosis*

Fig. 3. Data extracted from HIS database by wrapper agent



what kind of pain?

By rebgreen46 | Mar 07, 2015

I have where I cant walk very far my back hurts so bad I break out in a sweat. I cant stand up very long with out the pain. if I can sit down for a while it stops. and my legs and shoulders aches. is this my osteoporosis. I have bern diagnosed with it but the pain is getting worse. what is the best medication to take?

Tags: Pain, medication, hurts

Fig. 4. A post extracted from a forum webiste by wrapper agent

integration mechanisms for either structured or semi-structured data sources (or both) should be investigated, as for instance agent-based approaches, also in order to fully exploit the cooperation of agents we outlined in the main architecture to provide effective automation in the integration process

- finally, a significant further work is required to implement both user and disease agents and to explore how to leverage corresponding networks to improve integration and the quality of health-related information.

Acknowledgment

This work was supported by the "Programma Operativo Nazionale Ricerca & Competitivita' 2007–2013" within the project "PON04a2_C - Smart Health 2.0".

References

[1] K. J. Cios and W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, pp. 1–24, 2002.

[2] R. Haux, "Health information systems - past, present, future." *I. J. Medical Informatics*, vol. 75, no. 3-4, pp. 268–281, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/ijmi/ijmi75.html#Haux06

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/2047-2501-2-3

[4] J. Fisher and M. Clayton, "Who gives a tweet: Assessing patients interest in the use of social media for health care," *Worldviews on Evidence-Based Nursing*, vol. 9, no. 2, pp. 100–108, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1741-6787.2012.00243.x

[5] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health." in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#PaulD11

[6] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," *ACM Comput. Surv.*, vol. 18, no. 4, pp. 323–364, December 1986. [Online]. Available: http://doi.acm.org/10.1145/27633.27634

[7] A. Doan and A. Y. Halevy, "Semantic-integration research in the database community," *AI Mag.*, vol. 26, no. 1, pp. 83–94, March 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1090488.1090497

[8] J. F. Dipnall, M. Berk, F. N. Jacka, L. J. Williams, S. Dodd, and J. A. Pasco, "Data integration protocol in ten-steps (dipit): A new standard for medical researchers," *Methods*, vol. 69, no. 3, pp. 237 – 246, 2014, recent development in bioinformatics for utilizing omics data. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1046202314002382

[9] C. Heeter, "{EHR} progress and future outlook," {*AORN*} *Journal*, vol. 97, no. 3, pp. C7 – C8, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001209213001506

[10] S. Sachdeva and S. Bhalla, "Semantic interoperability in standardized electronic health record databases," *J. Data and Information Quality*, vol. 3, no. 1, pp. 1:1–1:37, May 2012. [Online]. Available: http://doi.acm.org/10.1145/2166788.2166789

[11] A. Sheth, S. Agrawal, J. Lathem, N. Oldham, H. Wingate, P. Yadav, and K. Gallagher, "Active semantic electronic medical record," in *The Semantic Web - ISWC 2006*, ser. Lecture Notes in Computer Science, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, Eds. Springer Berlin Heidelberg, 2006, vol. 4273, pp. 913–926. [Online]. Available: http://dx.doi.org/10.1007/11926078_66

[12] A. Baird, F. North, and T. S. Raghu, "Personal health records (phr) and the future of the physician-patient relationship," in *Proceedings of the 2011 iConference*, ser. iConference '11. New York, NY, USA: ACM, 2011, pp. 281–288. [Online]. Available: http://doi.acm.org/10.1145/1940761.1940800

[13] D. Lee, R. Cornet, F. Lau, and N. de Keizer, "A survey of snomed-ct implementations," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 87 – 96, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046412001530

[14] Health Level Seven international, *http://www.hl7.org/index.cfm.*

[15] J. Puustjärvi and L. Puustjärvi, "Ontology-based integration of clinical documents," in *Proceedings of the 14th International Conference on Information Integration and Web-based Applications &#38; Services*, ser. IIWAS '12. New York, NY, USA: ACM, 2012, pp. 342–347. [Online]. Available: http://doi.acm.org/10.1145/2428736.2428799

[16] D. Perez-Rey, V. Maojo, M. Garca-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Snchez, and A. Sousa, "Ontofusion: Ontology-based integration of genomic and clinical databases," *Computers in Biology and Medicine*, vol. 36, no. 78, pp. 712 – 730, 2006, special Issue on Medical Ontologies. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010482505000740

[17] J. H. D. Cho, P. Sondhi, C. Zhai, and B. R. Schatz, "Resolving healthcare forum posts via similar thread retrieval," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. BCB '14. New York, NY, USA: ACM, 2014, pp. 33–42. [Online]. Available: http://doi.acm.org/10.1145/2649387.2649399

[18] G. Eysenbach, "Infodemiology and Infoveillance," *American Journal of Preventive Medicine*, vol. 40, no. 5, pp. S154–S158, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.amepre.2011.02.006

[19] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, April 2011, pp. 702–707.

[20] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PLoS One*, vol. 6, no. 5, 2011.

[21] A. Longheu, V. Carchiolo, and M. Malgeri, "Personal health record feeding via medical forums," in *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 19th International Conference on.* IEEE, 2015. [Online]. Available: toappear

[22] A. Longheu, V. Carchiolo, and MicheleMalgeri, "Medical data integration with snomed-ct and hl7," in *New Contributions in Information Systems and Technologies*, ser. Advances in Intelligent Systems and Computing, A. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, Eds., vol. 353. Springer International Publishing, 2015, pp. 1165–1171. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16486-1_115

[23] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization, 2nd ed.* Amsterdam: John Benjamins, 2007.

[24] R. Baeza-yates and B. Ribeiro-Neto, *Modern Information Retrievial.* Seattle, Washington, United States: ACM Press, 1999.

[25] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093 – 1113, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2090447914000550

[26] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the First ACM Conference on Online Social Networks*, ser. COSN '13. New York, NY, USA: ACM, 2013, pp. 27–38. [Online]. Available: http://doi.acm.org/10.1145/2512938.2512951

[27] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, January 2008. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[28] Natural Language Toolkit chunk package, *http://www.nltk.org/api/nltk.chunk.html.*

[29] SNOMED CT, *http://www.ihtsdo.org/snomed-ct.*

[30] IHTSDO SNOMED CT Browser, *http://browser.ihtsdotools.org/.*

[31] A. Longheu, V. Carchiolo, and M. Malgeri, "Schema and data integration for relational and object-oriented data sources," in *Computer Science and Informatics (CSI2002), Proceedings of the 2002 Sixth International Conference on*, 2002.

[32] Fast Healthcare Interoperability Resources, http://hl7.org/implement/standards/fhir/.

[33] Microsoft HealthVault, https://www.healthvault.com/it/en-US.

[34] Patientslikeme, https://www.patientslikeme.com/.

[35] PHR - Medline, http://www.nlm.nih.gov/medlineplus/personalhealthrecords.html.

[36] OpenEMR, http://open-emr.org/.

[37] MedHelp, http://www.medhelp.org/.