# Semantic Knowledge Oriented Electronic Health Record Integration System

Winta Tesfaye
Tulane International Ethiopia, Addis Ababa, Ethiopia
wintes@gmail.com

Fekade Getahun
Department of Computer Science, Addis Ababa University, Ethiopia
fekadegetahun@gmail.com

**Abstract**

Electronic Health Record system aims to enhance the health care service. However, as the amount of information being stored grows, it not only makes the patient data duplication prone but also affects the quality of the health care service. This is so as a patient's medical history is distributed across several facilities and is not synced.

Duplicate records can happen because of many reasons - misspellings, typos, and transpositions. The high demand of data sharing especially in the domain of the medical world is highly related with providing a better health care and that possibly leads to less loss of life. This calls for data integration. In a country like Ethiopia, a nationwide unique patient identifier does not exist. Rather the concept of master patient index restricted to a heath facility is used. Hence, a patient identification number generating system is one of the first steps towards a unified view of a patient's medical history. In this paper, we have proposed an algorithm that helps us produce a unique patient identification number. Going further, when trying to merge or integrate heterogeneous and distributed information, there is a need to compute the similarity between patient profile and medical history. We identify similar patient through Edit Distance based similarity method and corresponding medical history matching using vector based similarity method - cosine with the help of dedicated knowledge base. In the last stage, we merge the clinical data of the patient as a result of these combined matchers.

*Keywords:* Semantic; Knowledge Oriented; Integration; Electronic Health Record System; Schema Matching; Patient Identification

## 1. Introduction

In recent years, there have been a significant amount of research that have focused on integration of data that are disparate and fragmented across different sources. It is the rapid growth in the amount of information that is being stored in such data sources that made the concept of data integration techniques a pressing issue and lead to this study.

The high demand of data sharing, especially in the domain of the medical world, is highly related with providing a better health care and that possibly leads to less loss of life. Healthcare organizations demand to unite and share the very sensitive patient data so as to increase the quality of care which in turn reduces treatment costs. Due to the abundance of great volumes of health records, there is a need to have a unified view of a patient's entire medical history. This need and the increasing awareness of added value through integration of medical information and to improve patient care is what has called for improvement in ways of integration.

## 2. Background

The first known medical record was developed in the fifth century B.C. with two basic goals [1]:

- a medical record should accurately reflect the course of disease, and
- a medical record should indicate the probable causes of disease.

These goals are still appropriate, but electronic health record systems can also provide additional functionality, such as interactive alerts to clinicians, interactive flow sheets, and tailored order sets, all of which cannot be done with paper-based systems.

The first EHRs (Electronic Health Record Systems) began to appear in the 1960s. In [2], it is reported that by 1965, at least 73 hospitals and

clinical information projects and 28 projects for storage and retrieval of medical documents and other clinically-relevant information were underway.

*Motivation Scenario*

Assume that Patient P1 goes to a health facility, HF1, to get health care service. In HF1, s/he gets a unique patient ID upon the first visit and registration process. The patient receives a medical diagnosis (for instance, Asthmatic) with or without a confirmed laboratory result and the medical practitioner may prescribe medicine and provide consultation geared to the patient case. Even though health record can be kept using paper or an electronic counterpart, this paper focuses on the electronic version.

After a few months, the same patient P1 goes to a different health facility, HF2. In HF2, P1 is given a different patient ID and is considered as someone who goes to a health facility for the first time or with a different prior set of primary disease diagnosis, as patient ID is restricted to one Health Facility. As a result of this, patients' data are scattered across facilities each having possibly different repository structure. Because of the lack of past patient history that accompanies each patient, in HF2, P1 was diagnosed for early stage of high blood pressure and given Aspirin wrongly.

This scenario demonstrated the need to

- identify patients independent of the health facility,
- match patients demographics and/or patient summary information based on the actual content, and
- integrate patients' records distributed in several locations.

## 3. The Proposed Solution

The aim of this work is to provide medical practitioners with a unified view of patient's medical history regardless of which facility the patient has received the service before. The high level architecture of our approach is shown in Figure 1. One of the main results of this work is generating nationwide patient ID. Unique patient ID is crucial to facilitate the integration process. Merging of patient information from different sources is assisted with element name mapper that identifies the elements/attributes that describe the same fact. The mapper uses our dedicated knowledge base that contains list of related concepts. Once the elements are mapped, we compute the content similarity using a text similarity module which is assisted with vector space mode. We have considered similarity matchers and their combined result to identify a certain patient across different databases. Then the clinical data of the patient shall be merged to be able to be viewed as one by a physician.

In order to briefly go through the details of the above overview, we have classified the report into three phases: the Patient Identification, Schema Matching, and Merger.

Among the different demographic information of a patient, the use of static and less frequently changing attributes is crucial as input to the ID generator. Thus, we propose the use of Birth-Place, Date-of-birth, Sex, and Blood type to uniquely identify each patient.

The algorithm that produces the output is shown in Algorithm 1.It returns a 14 digit value resulting from the concatenation of representing a Patient Uniquely as shown in Line 12.

We used edit distance to compare the similarity of the element names and content of the schemas. Edit Distance [9] is a string metric for measuring the amount of difference between two strings. The Edit distance between two strings is defined as the minimum cost sequence of operations to transform one string to the other. The basic edit script operations are: I (insert), D (delete), and R (Replace).

*Example:* Given two strings S1: 'Error' and S2: 'Eror', an insertion of r in S2 or removal of the double r from S1 makes the string similar, thus EditDistance(Error, Eror) = 1.

Similarly the EditDistance (great, grate) = 2

The higher the edit distance, the smaller is the similarity between the two strings. Thus, the edit distance similarity between two strings S1 and S2 is defined as

$$simedit(S_1, S_2) = \frac{1}{(1 + EditDistance(S_1, S_2))}$$

*Example:*

- $simedit(Error,Eror) = \frac{1}{(1 + EditDistance(Error,Eror))} = 0.5$
- $simedit(great,grate) = \frac{1}{(1 + EditDistance(great,grate))} = 0.33$

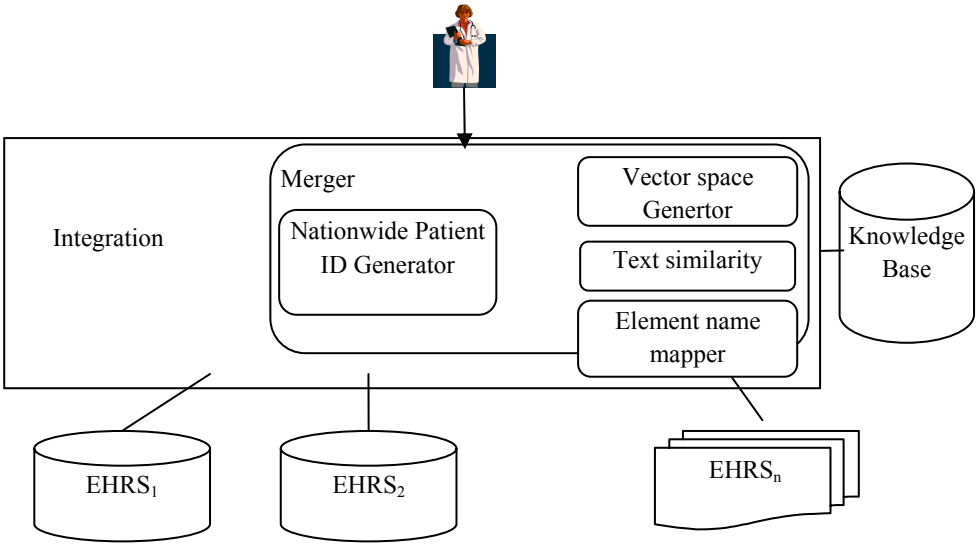


*Figure 1:* Overview of architecture of knowledge base integration of EHRS

*Algorithm 1*: Nationwide Patient ID Generator

| Line # | |
|---|---|
| | Input: |
| 1. | BirthPlace: string |
| 2. | Dateofbirth: date |
| 3. | Sex: char //sex as M or F |
| 4. | B: char // blood type |
| | Variables: |
| 5. | SSS: string, PPPPP, YYYMD, SSS: char [5] |
| 6. | KB: LocalityKnowledgeBase |
| | Output: |
| 7. | PatientID: Char[14] // PPPPP YYYMD B SSS |
| | Begin |
| 8. | PPPPP = GetLocalityCode(BirthPlace, KB) |
| 9. | YYYMD = GetDateBirthSex(Dateofbirth, Sex) |
| 10. | B: stands for Blood Type |
| 11. | SSS = GetNextSerialNo(PPPPP, DDDDD) //returns is a sequential number |
| 12. | Return PPPPP + DDDDD + B + SSS |
| | End |

The algorithm accepts the patient profile as input and visualizes to the user the relevant medical history. It first extracts the nationwide patient ID of the input using the function *ExtractNationWidePatientID*. Then it looks for the medical history of the patient with the given nationwide patient ID using the function *GetPatientMedicalHistorywithNationWidePatientID*. If the result is successful, merging comes down to applying conflict resolution also called merging rule presented here as function *ApplyConflictSolvingRules*. Otherwise, we look for list of patients profile very similar to the input profile. The similarity is computed using edit distance combined with vector based similarity – cosine. Notice that two profiles are similar if their similarity value is more than the threshold provided

by the user and returns the medical history of similar patients.

## 4. Experimentation (Prototype)

The prototype was developed based on the requirements at each phase of this work. The development tools used for the Semantic Knowledge Oriented Electronic Health Record Integration System are as follows:

Microsoft visual Studio 2005 for IDE Integrated development environment is used to parse through the database records as input and we used C# programming language. We used Microsoft SQL Server 2005 to keep our mediated schema that is used to show a unified view for the end user. Figure 2 shows the user interface used to register patient information which could be done in any of the health care facilities.

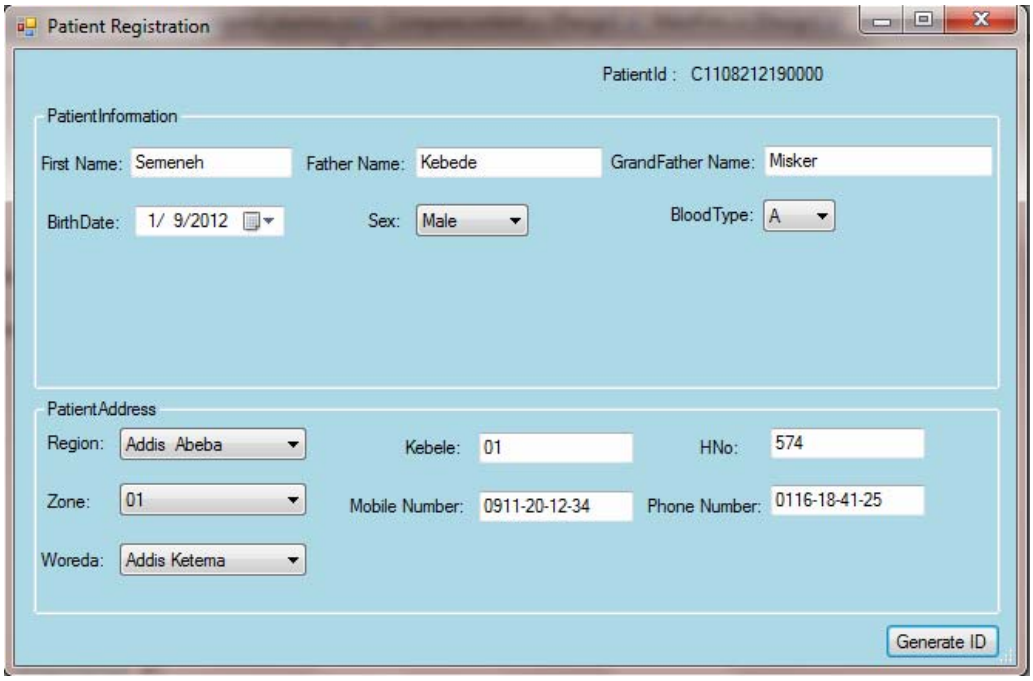


*Figure 1:* Sample Patient Profile with unique identification number

## 5. Related Work

### *5.1 Patient Identification number*

In the United Kingdom [3], the National Health Service (NHS) proposed National Unique Patient Identifier to be used in England and Wales. They are in the middle of implementation by applying a unique patient identification number throughout the health facilities on each patient that comes for a service. Each NHS Number is made up of 10 digits shown in a 3-3-4 format, the first 9 being the identifier and the last digit being a check digit.

According to the recent RAND Corp. study, in the United States [4], it is indicated that the use of a unique patient identification number for every person in the United States would help in reducing medical errors, simplify the use of electronic medical records, increase overall efficiency, and protect patient privacy.

### *5.2 Schema Matching*

CUPID is a hybrid match approach combining a name matcher with a structural one. Input schemas are converted to trees in which additional nodes are added to resolve the multiple relationships between a shared node and its parent nodes [5].

Similarity Flooding (SF) converts schemas (Relational, RDF, XML) into labeled graphs and uses fix-point computation to determine correspondences of 1:1 local and m:n global cardinality between corresponding nodes of the graphs [6, 7].

### *5.3 Similarity Computation*

Given a hierarchical knowledge base KB, collection of concepts related with is kind of relationship, Wu & Palmer evaluated a conceptual similarity between pair of concepts in hierarchy-based structure using the distance between their most common ancestor [8]. The similarity measure is denoted as:

$$sim(C_1, C_1) = \frac{2 * depth\ (C)}{(len(C_1, C) + len(C, C_1) + 2 * depth(C))}$$

where:-

- C is the least common ancestor that subsumes both $C_1$ and $C_2$
- depth(C) is the depth of C from the root
- len($C_1$, $C_2$) returns the path length between $C_1$ and $C_2$

*Example:* For instance, using the knowledge base shown in Figure 2, let us compute the semantic similarity between pair of concepts.

## 6. Conclusion

In this paper we have been able to provide approaches that provide unified patient identification number and merge/integrate heterogeneous patient data. Having an electronic health record system comes with a requirement of integration. We have seen the different experiences of developed countries such as the United States and the United Kingdom and learnt the pros and cons in order to recommend a separate identification number for patients. We have also seen the demand for integration of patient data. Studying through all the significant data matchers and merging data, we have provided a number of solutions. The contribution of this paper can be summarized as follows.

- an approach that generates a unified unique patient identification,
- a knowledge base schema matcher that results in the integration of the schema generated from the heterogeneous database schema of the electronic health record system,
- merging approach that puts together medical information of patients, and
- developed a prototype that contains the different algorithms developed in this work.

## References

[1] J. H. van Bemmel and M. A. Musen, ed., Handbook of Medical Informatics, Springer, The Netherlands, 1997, pp. 99.

[2] Richard Dick, Elaine B. Steen, and Don Detmer, editors, The Computer Based Patient Record: An Essential Technology for Health Care, Institute of Medicine, National Academy Press, 1997, pp. 111.

[3] The National Patient Safety Agency, NHS Number to be used as the unique patient identifier by all NHS organizations in England and Wales, September 2008.

[4] RAND Corporation, Creating Unique Health ID Numbers Would Improve Health Care Quality, Efficiency, Study Claims, October 2008.

[5] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm, Generic Schema Matching with Cupid, August 2001.

[6] Melnik, S., H. Garcia-Molina, and E. Rahm, "Similarity Flooding - A Versatile Graph Matching Algorithm", Proc. 18th Intl. Conf. Data Engineering (ICDE), San Jose CA, 2002.

[7] Amol Deshpande and Anthony Hunter, "Scalable Uncertainty Management" 4th International Conference, SUM 2010, Toulouse, France, September 27-29, 2010.

[8] Wu, Z. and M. S. Palmer, "Verb Semantics and Lexical Selection", In ACL., pp.133-138, 1994.

[9] Nick Koudas, Sunita Sarawagi, and Divesh Srivastava, "Record linkage: Similarity Measures and Algorithms", September 2006.