ASSURING INFORMATION QUALITY OF ELECTRONIC HEALTH RECORDS IN EHEALTH PLATFORM

(Research-in-Progress)

Ying Su
Information Quality Lab, Center for Resource Sharing Promotion
Institute of Scientific and Technical Information of China, Beijing, CHN
suy.rspc@istic.ac.cn

Ling Yin
Neural Information Centre, General Hospital of the People's Liberation Army
No. 28 Fu Xing Road, Beijing 100853, China
yinling301@126.com

Latif Al-Hakim
Department of Management and Marketing, Faculty of Business
University of Southern Queensland, Queensland 4350, Australia
hakim@usq.edu.au

Abstract: The core application of the eHealth is an electronic health record (EHR) system. One significant obstacle is the class of problems that arise due to variations in the quality of the information being shared. In this paper we outline a framework for assuring information quality (IQ) in the EHR context, using semiotics theory, semantic explanation of resources, and data couplings. Physicians can define the quality characteristics that are of importance in their particular domain by applying an IQ semiotics, which classifies and organizes these domain-specific quality characteristics within this quality assurance framework. Resource Description Framework (RDF) is used to explain data resources, with reference to IQ indicators defined in the semiotics. Data couplings - again defined in RDF - are used to represent mappings between data elements and the IQ semiotics. As a practical illustration of our approach, we present a case study from an open eHealth platform for the community-wide health information network of China.

Key Words: Information quality; electronic health record; EHR; eHealth Resource Description Framework;

1. INTRODUCTION

Information is viewed as a fundamental resource in the discovery of new scientific knowledge. Physicians expect to make use of information produced by other labs and projects in validating and interpreting their own results. A key element of eHealth is the development of a stable environment for the conduct of information-intensive forms of science. Problems arise due to variations in the quality of the information being shared [1]. Data sets that are incomplete, inconsistent, or inaccurate can still be useful when physicians are aware of these deficiencies.

Conceptual frameworks of information quality abound in management, communication, and information technology literature. In our review of information quality literature from the last ten years, we have found thirty information quality frameworks that define and categorize quality criterion for information in various application contexts, such as media studies[2, 3], data warehouses[4], information system[5], or knowledge management[6]. We are developing techniques for assuring information quality (IQ) using Semantic Web. In contrast to previous IQ research, which has tended to focus on the identification of

generic, domain-independent quality characteristics (such as accuracy, currency and completeness) [7], we allow physicians to define the quality characteristics that are of importance in their particular domain. For example, one group of physicians may define “accuracy” in terms of some calculated experimental error, while others might use it as a standard of the type of equipment that captured the data.

In order to support this form of domain-specific IQ, we identify three key requirements, each of which can be met using Semantic Web:

- Physicians can not only use the domain-specific IQ descriptions but also to reuse definitions created by others. To meet this requirement, we propose an extensible IQ semiotics containing basic domain-independent IQ terms, upon which definitions of domain-specific concepts can be built.
- IQ descriptions for specific resources need to be computed and associated with those resources. This can be done by attaching origin information to the RDF explanation instances.
- Resources include data and services; both of these kinds of resource are modeled by concepts in the IQ semiotics, so that the semiotics can express which kinds of IQ descriptor make sense for which kinds of resource. We refer to these relationships as *couplings*, which can be captured using an RDF schema.

The rest of the article is organized as follows. Section 2 gives an overview of the IQ Assurance (IQA) framework, and the following sections present each of the three components in detail: Section 3 introduces the IQ semiotics, Section 4 describes the coupling schema, and Section 5 presents the explanation model. We conclude in Section 6 with lessons learned and future work.

2. AN IQ ASSURANCE FRAMEWORK

The following framework strives to fulfill most of above three requirements [8]. This framework consists of four major elements: The first element is the framework's vertical structure. It consists of four sorts of information quality that categorize crucial information according to semiotics theory[9]. Semiotics are the sign’s actual representation; its referent or intended meaning (i.e. the phenomenon being represented); and its interpretation or received meaning (i.e. the effect of the representation on an interpreter’s actions, that is, the actual use of the representation). Informally, these three components can be described as the form, meaning, and use of a sign. Relations between these three aspects of a sign were further described by Morris as syntactic (between sign representations), semantic (between a representation and its referent), and pragmatic (between the representation and the interpretation) semiotic levels. Again, informally, these three levels can be said to pertain to the form, meaning, and use of a sign respectively. With this background, the correspondence between semiotics and information quality can be clarified and the applicability of semiotics to the formal definition of information quality justified[10].

The next element of the framework is the horizontal structure, which is divided into four phases. The four phases represent the mechanism to assure information quality from physicians’ point of view. The IQ criterion, — the third component of the framework — is placed along these phases according to their importance for the different phases. How IQ criterion can be applied is answered by the last element of the framework, the assuring principles. The principles help to assure the quality of information in every phase. Figure 1 sets out the key relationships between the various individuals and classes.

IQ semiotics includes definitions of domain-independent IQ concepts and classes of domain-specific indicator. It also models the various kinds of abstract data entities to which we might wish to apply IQ indicators. Finally, the semiotics defines the various kinds of data assuring function available, as described in detail in Section 3.

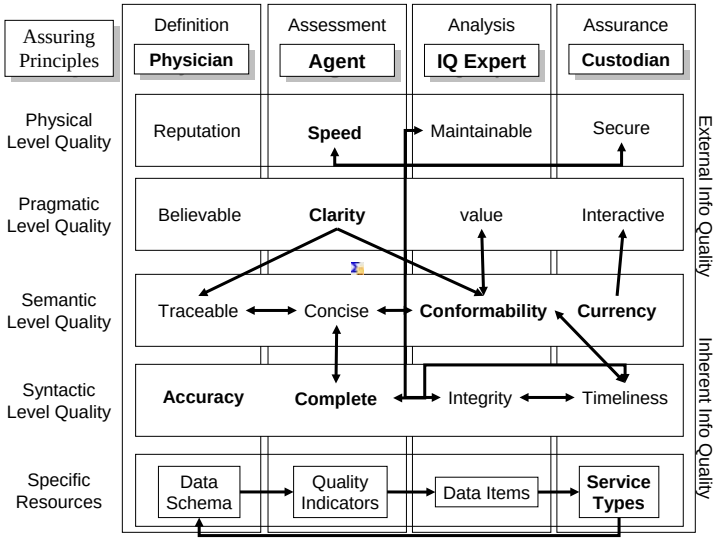

Figure 1 Overview of the elements of the IQA framework

At the bottom of Figure 1 we have instances of specific resources, which are represented in XML. Couplings and explanations both relate resources to elements of the IQ semiotics. An instance of a coupling relates a resource instance to the corresponding class in the semiotics. One of the main uses of couplings is to determine which parts of the IQ conceptualization are relevant to a particular concrete data model. For details, see Section 4. An instance of an explanation relates a specific resource instance to the instance of a quality concept. The IQ instance is said to explain the associated resource. Further details of explanations are given in Section 5. The difference between couplings and explanations is that the former relate data schema elements (or service types) to the corresponding semiotics classes, while the latter relate individual items of data to individual pieces of quality evidence. In other words, couplings define which IQ concepts relate to which kinds of data or service, while explanations associate individual computed IQ descriptors with specific pieces of data.

3. A SEMIOTICS MODEL FOR IQ ASSURANCE

EHRs were classified on the basis of the International Organization for Standardization (ISO) definition. According to this definition, the EHR means a repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information and its primary purpose is to support continuing, efficient and quality integrated healthcare. The physicians' exercise is one of knowledge elicitation: the tacit knowledge regarding quality properties of interest needs to be made explicit and formalized. We now present a semiotics model that supports such a knowledge elicitation process, by providing a vocabulary and semantic structure for assuring information quality. The model allows physicians to share and reuse their understanding of quality, as well as to perform semi-automated quality assessments on data sets of interest to them.

3.1 Basic Semiotics Structure

The structure and content of EHRs has varied over time, and we made a distinction between time-oriented, problem-oriented and source-oriented EHRs. Nowadays EHRs combine all three elements. Our model is based on the assumption that physicians should not be concerned with abstract IQ definitions, such as Currency, Completeness or Accuracy, and that they should instead be able to state their quality requirements in operational terms, by describing decision procedures that determine the suitability of the

data.

In the semiotics, we model IQ concepts by introducing **Quality Assurances** (**QA**); these are decision procedures that are based upon some **Quality Evidence** (**QE**), which consists either of measurable attributes called **Quality Indicators**, or recursively, of functions of those indicators, **Quality Metrics**. Three main sources of indicators are common in practice:

- Origin metadata, which provides a description of the processes that were involved in producing the data.
- Quality functions that explicitly measure some quality property, these functions are typically available from toolkits for data quality assessment with reference to specific issues.
- Metadata that is produced as part of the data processing.

Focusing primarily on the second and third category, we model the indicator-bearing environment as a collection of **Data Analysis Tools** that may incorporate multiple **Data Calculation functions**, and which are applied to some **Data Entity**. Indicators are either parameters to or output of these analysis tools. A **QA** is applied to collections of data items, which are individuals of the **Data Entity** class, using the values for the indicators associated to those items. The practical quality metrics are part of the output of a calculation function called **QMCalculator**, used in the **IQA Calculator Analysis Tool**. A quality metric called **IQA Calculator Ranking** associates a score to each data in the set, using a function of indicators. This score can be used either to classify data as acceptable/non acceptable according to a user-defined threshold, or to rank the data set. Here we will assume that our decision procedure is a grade function called **QA-Func**, that provides a simple binary grade of the data set according to the credibility score and to a user-defined threshold.

The following is a summary of the classes and relationships introduced above, using informal notation for the sake of readability; user-defined axioms.

1. **Quality-Assurance** is based on **Quality-Evidence**;
2. **Quality-Indicator** is-a **Quality-Evidence**;
3. **Quality-Metric** is-a **Quality-Evidence**;
4. **Quality-Metric** is based on **Quality-Indicator**;
5. **Quality-Evidence** is output of **Data-Calculate-function**;
6. **Data-analysis**-tool is based on **Data-Calculate-function**;

3.2 IQ Calculation Function

The main idea for semiotics model is to encourage physicians to explain their domain-specific concepts with simple and concrete quality features, to the extent that they are familiar with them, and to use reasoning over Web Ontology Language Description Logic (OWL DL) to entail additional quality properties, or to determine inconsistencies.

Building on the structure described so far, we begin by adding a top-level Quality **Property** class, with a number of subclasses for **Consistency**, **Timeliness**, **Currency**, and more. Our collection for these concepts currently includes about 20 classes, organized into a four-level hierarchy. Also, we add a root

class for **Quality Criterion**, whose subclasses include **Syntactic-QC**, **Semantic-QC**, **Pragmatic-QC**, and **Physical-QC**. These are examples of the "concrete" properties that physicians can more easily associate to specific indicators, or indicator-bearing functions or tools. Thus, we expect hospitals to be able to assert that the **QMCalculator** function has a **Syntactic-QC**, because its purpose, from the quality perspective, is to provide information on the syntactic in the experiment result. Note that the semiotics model allows a single piece of evidence, or function, to have multiple quality criterions. The only user assertion for the example is:

QMCalculatorReport has quality characterization **Syntactic-QC**. We then introduce OWL DL axioms that describe classes of evidence that have the same quality criterion; given that hospitals may quality-characterize either indicators, metrics, functions, or tools, a sample definition is as follows:

Syntactic evidence includes all and only the quality metrics or indicators whose quality criterion includes **Syntactic-QC**, union all indicators that are output of functions, or of tools that use functions, whose quality criterion includes **Syntactic-QC**. Here is the OWL DL definition for this class:

SyntacticEvidence $\equiv$ (QtyMetric $\cap$ ($\exists$ metric-based-on-indicator SyntacticEvidence) $\cup$ (QtyIndicator $\cap$ $\exists$ is-output-of ($\exists$ has QC SyntacticQC)) $\cup$ (QtyIndicator $\cap$ $\exists$ is-parameter-of ($\exists$ has QC SyntacticQC)) $\cup$ (QtyIndicator $\cap$ $\exists$ has QC SyntacticQC)

Using the user-defined assertion above, the definitions in the previous section, and this class definition, an OWL DL reasoner entails the following:

- QMCalculatorReport $\subseteq$ SyntacticEvidence,
- AccuRate $\subseteq$ SyntacticEvidence,
- MemberRatio $\subseteq$ SyntacticEvidence,
- IQMCalculatorRanking $\subseteq$ SyntacticEvidence.

We now define the Accuracy class in terms of the underlying quality criterion, expressing the following:

Any quality property that is based on a decision procedure that makes use of Syntactic or Pragmatic evidence, can be classified as Accuracy. Formally:

Accuracy $\equiv$ ($\exists$ QtyProperty-from-QtyPreference ($\exists$ pref-based-on-evidence (SyntacticEvidence PragmaticEvidence))

This last definition allows the semiotics to be extended in a consistent way using standard reasoning. Firstly, given a user-defined but yet unclassified quality property, let us call it QA-Satisfaction, that is based on the QA-Func procedure, the reasoner entails that the property is a subclass of Accuracy. Conversely, hospitals may classify QA-Satisfaction within the IQ top-level taxonomy; in this case, the reasoner verifies the consistency of this assurance.

4. A COUPLING MODEL FOR IQ ASSURANCE

As we have shown, the IQ semiotics includes semiotics models of data resources and the quality assurance services which can be applied to them. The actual data resources have a native definition and presentations; quality calculation functions applicable on the data might have multiple implementations in different programming languages. We designed a generic data model to capture the mapping relationships between data or service resources and their semantic definition. The basic structure of the coupling model is presented in Figure 2. There are four core concepts in the Coupling model:

Resource refers to any resource that can be located on the EHR. We distinguish two sub types of resource: **DataResource** and **ServiceResource**. The former refers to any resource which stores information (e.g. an **XML** file or database table); the latter type represents any service, application or procedure which performs action on a **DataResource** (e.g. a Web service). We define three categories of **DataResource**:

- **DataEntityResource** represents elements defined in a data schema/structure, or a column defined in a **DB** table.
- **DataElementResource** represents a data element inside a collection, for example an XML element specified by an XPath, or a database tuple.
- **DataCollectionResource** represents a collection of data elements, for example an XML document, a database table, or a text file.

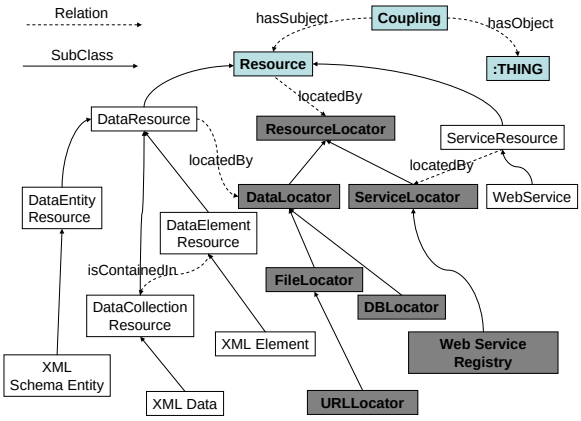

Figure 2 Overview of the IQA coupling model

A **Coupling** relates a **Resource** to a semantic mode in IQ semiotics structure (see Section 3.1). This relationship is defined by two properties on the coupling:

hasSubject identifies the subject of the coupling, which is always a locatable Resource.

hasObject identifies the object of the coupling, which can be any semantic concept in any semiotics (represented in our semiotics diagram with the most general concept **:THING** — for example, this could be any class in our IQ Semiotics).

ResourceLocator identifies a global locator for a specific resource. Since the resource is categorized into **DataResource** and **ServiceResource**, the **ResourceLocator** has two types: **DataLocator** and **ServiceLocator**. Due to various ways to access the data resources, the data locator can have different types. For example, for a data document, we can use a URL to retrieve it; while for a DB table, a DB connector API could be used (such as ODBC). Similarly, **ServiceLocator** has different types; for example, the locator of a quality annotation web service can be referred to a WSDL description and the endpoint of the service is presented in this WSDL description.

Couplings are bi-directional: the coupling from resource to concept is used to identify which IQ indicators and associated checking functions are applicable to a particular concrete (e.g. XML) data model; the coupling from concept to resource is used to locate concrete data and service implementations (e.g. Web services) to run a data check.

5. AN EXPLANATION MODEL FOR IQA

An important aspect of the IQ assurance is to share and reuse high quality information on data resources among user-physicians. In order to achieve this we provide a data model which formalizes explanation information with semantic support. The structure of our explanation model is shown in Figure 3. The concepts shaded in the figure are defined externally to the explanation model: prefixes c and s identify the coupling model and the IQ semiotics respectively.

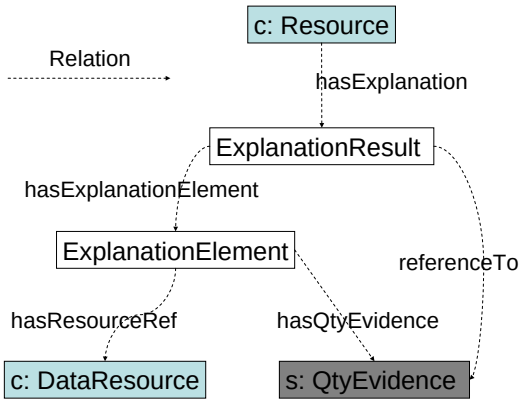

Figure 3 Structure of Explanation Model

The property **hasExplanation** represents the relationship that a ***c:* Resource** is explained with quality information recorded in an **ExplanationResult**. **ExplanationResult** defines a class of resource that records the output and related information from one run of some quality explanation service. These explanation results are a group of instances of one particular **s: QtyEvidence** class; the property **referenceTo** records the name of the relevant class, and the property **hasExplanationElement** records individual explanation result elements, each of which contains an individual **s: QtyEvidence** instance.

An **ExplanationElement** relates one individual instance of **s: QtyEvidence** to one individual explaind resource, using the properties **hasQtyEvidence** and **hasResourceRef** respectively.

The main advantage of this approach is flexibility: an explanation can be easily attached to any kind of resource, and easily associated with any IQ semiotics concept. We also support the attachment of origin information to instances of **ExplanationResult**, including the identification of the particular checking function used to generate the explanations, and the data selections used as input. Details of this origin information are omitted for space reasons; however, we are exploring the use of existing origin architectures for capturing some of these data.

6. CONCLUSION

The IQA framework offers a methodology for assuring information quality in an eHealth context, allowing user-physicians to specify their IQ requirements against a formal semiotics, so that the definitions are machine-manipulable. To the best of our knowledge, this semiotics is the first systematic attempt to capture generic and domain-dependent quality descriptors in a semiotics model. In this paper, we have shown how the use of OWL DL supports extensibility of the core semiotics with domain-specific quality definitions. We have also introduced coupling and explanation models that serve to associate concepts in the IQ semiotics with data and service entities. Couplings allow IQ-aware tools to identify parts of the IQ semiotics relevant to a specific data model. Explanations attach quality metadata to resources. Both the coupling and explanation models are to some extent intended to be generic, reusable components.

The IQA framework has been implemented in a collection of services accessible from a physician's desktop environment. We are currently gathering feedback from our collaborating hospitals, after which we aim to further develop the IQA framework and associated toolset.

ACKNOWLEDGEMENT

We would like to thank NNSFC (National Natural Science Foundation of China) for supporting Ying Su with a project (70772021, 70831003).

REFERENCES

[1] L. P. English, *Improving data warehouse and business information quality methods for reducing costs and increasing profits* New York: Wiley, 1999.

[2] M. Lang, S. Southard, and C. Bates, "Group Performance and Collaborative Technology: A Longitudinal and Multilevel Analysis of Information Quality, Contribution Equity, and Members' Satisfaction in Computer-mediated Groups," *Technical Communication,* vol. 53, pp. 271-271, 2006.

[3] W. L. Bennett, "Global media and politics: Transnational communication regimes and civic cultures," *Annual Review of Political Science,* vol. 7, pp. 125-148, 2004.

[4] R. Busatto, "Using Time Series to Assess Data Quality in Telecommunications Data Warehouses," in *International Conference on Information Quality (MIT IQ Conference)*, MIT, 2000, p. 10.

[5] C. Cappiello, C. Francalanci, and B. Pernici, "Time-related factors of data quality in multichannel information systems," *Journal of Management Information Systems,* vol. 20, pp. 71-91, Win 2003.

[6] H. Jerssica, H. Kuan-Tsae, J. K. Kuse, S. Geng-Wen, and W. Ko-Yang, "Customer Information Quality and Knowledge Management: A Case Study Using Knowledge Cockpit," *Journal of Knowledge Management,* vol. 1, pp. 225-236, 1997.

[7] L. Al-Hakim, "Information quality management: theory and applications," Hershey, PA : Idea Group Pub., 2006, p. 301.

[8] Y. Su and Z. Jin, "Assuring Information Quality in Knowledge intensive business services," in *3rd International Conference on Wireless Communications, Networking, and Mobile Computing (WiCOM '07)*, Shanghai, China, 2007, pp. 3243-3246.

[9] R. S. Corrington, *A semiotic theory of theology and philosophy*. Cambridge, UK ; New York, NY Cambridge University Press, 2000.

[10] R. Price and G. Shanks, "A semiotic information quality framework: development and comparative analysis," *Journal of Information Technology*, vol. 20, pp. 88-102, Jun 2005.