Experiences with a Two-Level Modelling Approach to Electronic Health Records

L. Bird, A. Goodchild and Z. Tun

CRC for Enterprise Distributed Systems Technology
DSTC Pty Ltd, Level 7 General Purpose South
The University of Queensland, QLD 4072
Email: {bird, andrewg, zarzar}@dstc.edu.au

In today's distributed healthcare environment, information is a key asset and getting access to that data is vital to the management of a patient's health. Electronic Health Record (EHR) standards aim to assist in the interoperable access and integration of this distributed information. However, there are difficulties applying traditional system development approaches to EHR standards because of the complexity and range of clinical data. The Good Electronic Health Record (GEHR), a framework for structuring, storing and sharing EHRs, uses an innovative two-level modelling approach to overcome these difficulties. In this approach, all types of clinical data are stored using a generic model, and constraint models called "archetypes" ensure that the data represents valid clinical concepts. To test this approach, two field trials funded by the Australian General Practice Computing Group (GPCG) looked at exporting data from existing systems to GEHR. In this paper, we describe GEHR's "archetype" approach and describe the outcomes of the trials.

Cross References: H.1 – Information systems: Models and principles; H.4 – Information systems application; J.3 – Computer applications: Life and medical sciences

1. INTRODUCTION

As computerised systems become more prevalent in health care facilities to store and manage information about patients' health, exchanging data between these systems is important to provide better, coordinated care. However, health care facilities use a variety of information systems to manage patients' health data. In Australia, there are several practice health record systems, such as Medical Director[1] and Medical Spectrum[2], while hospitals also use a number of clinical information systems, such as CERNER[3] and iSoft[4]. Each of these systems varies in its underlying technology and approach for structuring health records. In order to be able to exchange information between these systems, without building a one-to-one gateway between every single vendor combination, a single common approach for structuring health records needs to be adopted.

---

[1] http://www.md.com.au (accessed 29 April 2003)
[2] http://www.medispec.com.au (accessed 29 April 2003)
[3] http://cerner.com (accessed 29 April 2003)
[4] http://www.isoft.com (accessed 29 April 2003)

There are many possible approaches to this problem, including using standards such as HL7 (Health Level 7)[5], CEN 13606[6] and CorbaMED[7]. However, a common problem between these approaches is that they do not provide a simple future-proof solution for standardising the ever increasing variety of clinical information structures such as clinical tests, notes, and care plans.

The Good Electronic Health Record (GEHR), discussed in section 2.4, is one approach that addresses this issue. GEHR uses a two level modelling approach, where all information is described using a generic health record model that enables a wide variety of health information to be stored and then the structure of that information is further constrained by an archetype. An archetype is a constraint model that limits the structure of certain kinds of information such as clinical tests, notes, care plans, etc. Thus the archetype can be used to ensure that only data of a certain structure and hence quality can be added to the record. Furthermore, as the archetypes are decoupled from the underlying health record model, new archetypes can be added over time allowing a health record system to evolve without substantial changes.

In this paper, we describe the latest Australian GEHR research, and present the outcomes of two Australian General Practice Computer Group funded trials to test the approach taken by GEHR. We begin, in Section 2, by defining electronic health records and comparing a number of approaches to building EHR systems, including the GEHR approach. In Section 3, we describe the implementation of two recent Australian GEHR trials, which investigated the process of exporting clinical information from existing systems into a GEHR-based repository. The lessons learnt during these two trials are discussed in Section 4, before conclusions are reached in Section 5.

2. ELECTRONIC HEALTH RECORDS AND GEHR

2.1 What is an Electronic Health Record (EHR)?

For the purposes of this paper, we will use the definition of an Electronic Health Record (EHR) from the Health Information Network for Australia (HINA) report (NEHRT, 2000), which states that an EHR is:

> *"An electronic longitudinal collection of personal health information, usually based on the individual, entered or accepted by health care providers, which can be distributed over a number of sites or aggregated at a particular source. The information is organized primarily to support continuing, efficient and quality health care. The record is under the control of the consumer and is stored and transmitted securely."*

The Electronic Health Record (EHR) is an important component for information management in an integrated healthcare system. The primary purpose of the EHR is to provide a documented record of care to be used as a means of communication among healthcare agents contributing to the consumer's care. This information can also be interpreted by automated decision support systems, which may provide alerts and advice to healthcare agents.

The kind of information accumulated in an EHR includes (Rector, 1992):

* *Retrospective:* a historical view of health status and interventions, such as test results, progress notes, referrals, orders, family history, past medications;
* *Concurrent:* a "now" view of health status and active interventions, such as current medications, current problems, therapeutic precautions, lifestyle; and
* *Prospective:* a future view of a patient's care, such as care plans, goals and targets.

---

[5] http://www.hl7.org (accessed 29 April 2003)

[6] http://www.centc251.org (accessed 29 April 2003)

[7] http://cgi.omg.org/corbamed (accessed 29 April 2003)

The primary beneficiaries of the shared EHR are the health care consumer and healthcare agents. However, it can also be put to secondary uses (ISO, 2002), such as:

- *Medico-legal purposes* – as evidence of the care provided, an indication of compliance with legislation or a reflection of the competence of the clinicians;
- *Quality management* – for continuous quality improvement studies, utilisation reviews, performance monitoring (peer review, clinical audits and outcomes analysis), benchmarking and accreditation;
- *Education* – where de-identified sample data can be used as case studies for teaching purposes;
- *Research* – for development and evaluation of new diagnostic modalities, disease prevention measures and treatments, epidemiological studies, population health analysis;
- *Policy development/health service management* – for health statistics analysis, trends analysis, casemix analysis, resource allocation, reports and publications, marketing strategies and enterprise risk management; and
- *Billing/finance/reimbursement* – for insurers, government agencies, funding bodies.

There are two basic kinds of EHR: a "shared EHR" and a "local EHR". It is expected that the "shared EHR" (i.e. the EHR that is shared between multiple healthcare organisations) will be used as a secondary source of information used to enhance communication between a group of disparate clinics, while each provider will typically consult a "local EHR" as a primary source of information within a local clinic. The shared EHR typically contains summarised information that is of interest to multiple types of providers, whereas a local EHR contains detailed information usually of interest to a single type of provider. For example, a podiatry clinic might keep detailed information about patients' feet, but share only information relevant to other providers, such as indication of diabetes in patients' feet.

2.2 Benefits and Challenges of Building a Common EHR Model

The electronic interchange of clinical information can solve a number of problems in healthcare. For example:

- Between 44,000 and 98,000 people die each year in the United States due to medical errors, where many of these deaths are attributed to severe adverse drug reactions (Kohn and Corrigan, 2000). Communicating vital information like adverse drug reaction histories in a timely fashion can prevent deaths and other serious consequences. A study into medical errors in Australian general practices found that communication problems, such as not informing GPs of the outcomes of hospital referrals and tests, were a major contributing factor to medical errors (Bhasale *et al*, 1998).
- Time is wasted when the same questions are repeatedly asked to obtain a patient's history. If a patient's history is shared electronically, not only will time be saved, but the quality of the data is likely to be better. Bhasale et al's study in Australian general practices (1998) also identifies inadequate recording of patients' history as another communication problem that contributes to medical errors.
- There is unnecessary duplication of tests (e.g. pathology and radiology tests) because health care providers may not have easy access to patients' previous test results. The HINA report (NEHRT, 2000, p. 171) estimates that AUD$56 million per year can be saved from avoiding duplications of tests in Australia with the introduction of an electronic network for exchanging health information.
- Delivering health care to a minority of chronically ill patient accounts for the majority of health care costs and often involves a great deal of collaboration between many professionals at

multiple different points of care. Electronic communications make it easier to disseminate timely information to all involved parties. A case study that looked into the effect of using electronic data exchange in a coordinated care environment, namely a diabetes clinic, found that communication between health care providers increased, they had better access to data, e.g. test results, and even a small improvement in patients' health was recorded over the short period (Branger *et al*, 1999).

However, there are several challenges to building a common EHR model to enable the exchange of clinical data. EHRs are complex, just as the human biology is complex. The EHR itself is not a single patient record in the health care domain. The data that constitutes an EHR comes from numerous records, such as nurse's notes, progress notes, care plans, test results, and many others. There is a wide range in complexity of this data from the highly unstructured, e.g. progress notes, to the semi-structured, e.g. discharge referrals and care plans, to the highly structured, e.g. biochemistry test results, multi-lead ECG (Electrocardiography), and MRI (Magnetic Resonance Imaging). At the same time, there is a large body of medical knowledge, which is constantly expanding. For example, SNOMED (Systemised Nomenclature of Medicine), which is a terminology set that categorises medical concepts, has over 350,000 terms. There are hundreds of clinical pathways to manage various conditions. In addition to new clinical concepts being introduced, existing ones change due to improvements in knowledge and processes. A common EHR model needs to be flexible to handle the complete range of complexities in data, be able to encompass a multitude of clinical concepts, and be adaptable to changes in the medical domain. These requirements make it difficult to build a common EHR model.

2.3 Approaches to Building EHR Systems

There are many approaches to building traditional EHR systems and standards. They can be categorised in the following ways:

* *Unstructured approach:* In this approach, the EHR is simply a warehouse filled with unstructured text. Figure 1 shows how a blood pressure instance might be represented. This approach allows an EHR system that can accommodate different forms of medical data as well as handle changes in the medical domain to be built rapidly. However, this system is of limited value in the long run because detailed data cannot be successfully queried and reported on for management or epidemiological purposes nor can it be used by decision support systems to inform health care providers of potential adverse effects of proposed therapies. An example of such a record architecture would be a text database containing HL7 CDA (Clinical Document Architecture) level 1 documents (Alschuler, 2000; Dolin *et al*, 2001).

bp: 120/80

Figure 1: Blood Pressure Instance in the Unstructured Approach

* *BIG model approach:* In this approach, the EHR is built by having a separate table or class for each clinical concept. Figure 2 shows an instance of the blood pressure class. These systems tend to have very large schemas, which, if printed out, could fill a wall. This, in turn, leads to errors in the systems because few people completely understand the entire model. Furthermore, the model becomes brittle over time as new concepts are introduced and existing ones are changed in the health care domain. An example of this record architecture is the one that is implicit in most current clinical GP and hospital software, and indeed most information systems today.

bp:BLOOD_PRESSURE
systolic = 120
diastolic = 80

Figure 2: Blood Pressure Instance in the Big Model Approach

- *Generic model approach:* In this approach, a generic model is designed to allow a wide variety of data to be accommodated in a general-purpose set of data structures. For example, instead of having a specific data structure devoted to biochemistry results and another to blood pressure readings, a single general-purpose structure is created for storing observations. Figure 3 shows an example of a blood pressure instance in this approach. The advantage of this approach is that the model is small enough to understand; yet many kinds of information can easily be stored in the generic structures. However, it is only a small improvement on the unstructured approach. Anything can be stored in these general-purpose structures leading to lower quality data, which in turn makes it difficult to query the data or use in decision support systems. Examples of such a record architecture include the original GEHR (Good European Health Record) and CEN 13606 models.
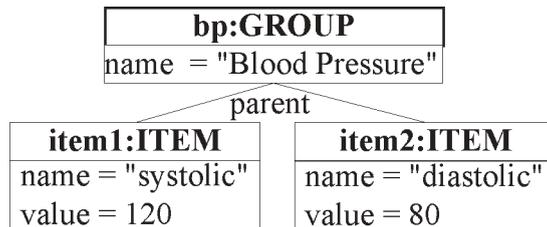
bp:GROUP
name = "Blood Pressure"

parent

item1:ITEM
name = "systolic"
value = 120

item2:ITEM
name = "diastolic"
value = 80

Figure 3: Blood Pressure Instance in the Generic Model Approach

Each of these approaches described have serious weaknesses. The unstructured and generic model approaches lead to lower quality and less 'useable' data, while the big model approach makes it difficult to manage the system over time. The Australian GEHR project introduced a new approach, discussed in the next section, which overcomes the weaknesses in the approaches described above.

2.4 GEHR

2.4.1 A Brief History of GEHR

GEHR started out as the Good European Health Record, a research project funded by AIM (Advanced Informatics in Medicine), a European Commission research initiative (GEHR, 2000). The goal of the project, which ran from 1992 to 1995, was to develop a common EHR architecture for Europe. Research into the EHR architecture developed by the GEHR project continued beyond its original grant. The project then took on a more international flavour with some key contributors moving to Australia, while interest appeared in non-European countries. These factors led to the renaming of the project to the Good Electronic Health Record.

The original GEHR project used the generic model approach, described in the previous section, since a variety of clinical data can be stored using a relatively small and manageable model. However, the project contributors realised that the data stored need not be clinically valid. This led

to the development of "archetype" concept by the Australian GEHR contingent, which is discussed next.

2.4.2 The Australian GEHR Approach
The approach developed by the Australian GEHR project builds upon the generic model approach. In this new approach, there is a generic model called a "reference model", which specifies how to:
* Organise and group clinical information
* Capture contextual information
* Query and update the health record
* Use versioning and attestation to safely manage clinical information from a medico-legal point of view.

Although the reference model has rich capabilities, it is generic enough to store any type of clinical information. To overcome the problem of lower data quality that results from using general-purpose structures, the Australian GEHR project introduced a constraint mechanism called "archetypes", which ensures that the stored information is valid in terms of clinical knowledge. Figure 4 below illustrates the relationship between the reference model, the archetype model, and instances of each.
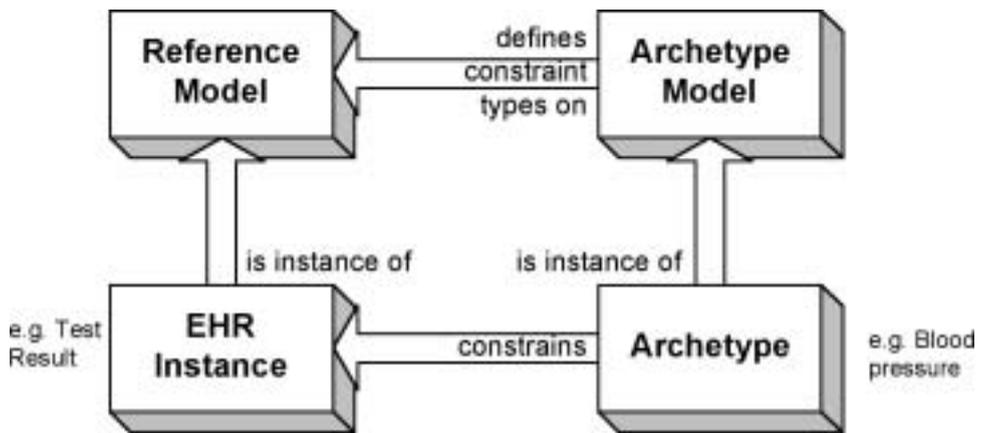

Figure 4: The relationship between models and instances in the GEHR approach

2.4.3 Archetypes
An archetype represents a single clinical concept, defining the content and shape of data that needs to be stored for that concept. In technical terms, an archetype specifies constraints on the data structures in the reference model.

As a simple example, a blood pressure archetype used to constrain the generic model shown in Figure 3 would specify that the first ITEM in a blood pressure GROUP will have the name, "systolic", and its value must be an integer between 40 and 300, while the second ITEM will have the name, "diastolic", with the same constraints on its value. This simple archetype is illustrated in Figure 5(b). Figure 5(a) shows the corresponding reference model for this archetype.

Note that, in practice, much richer constraint models can be used, as shown below in Figure 6.
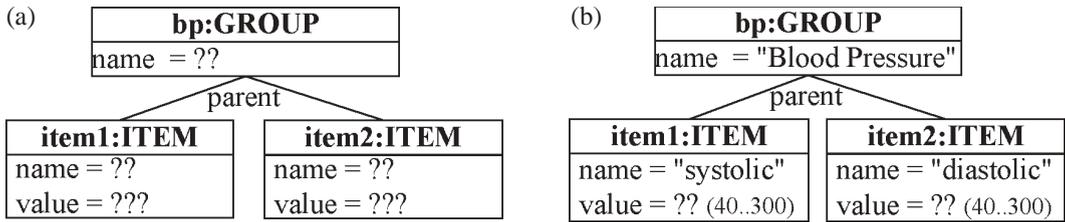
(a)

bp:GROUP
name = ??

parent

item1:ITEM
name = ??
value = ???

item2:ITEM
name = ??
value = ???

(b)

bp:GROUP
name = "Blood Pressure"

parent

item1:ITEM
name = "systolic"
value = ?? (40..300)

item2:ITEM
name = "diastolic"
value = ?? (40..300)

Figure 5: (a) Reference model and (b) archetype – both for Figure 3 instance

In Figure 6, the major classes in the tree (e.g. 'Observation Content', 'Proposition', 'Group' and 'Stat Ref') and the attribute types (e.g. 'name', 'recorder') are part of the fixed reference model, while the included attribute values (e.g. "blood pressure" and "systolic") and constraints (e.g. "mmHg") are part of the archetype constraint model.

One metaphor for understanding archetypes is that of language. The reference model is like a grammar for the English language and controlled vocabularies, such as SNOMED, are like the English dictionary. However, a vocabulary plus a grammar still allows you to write nonsense
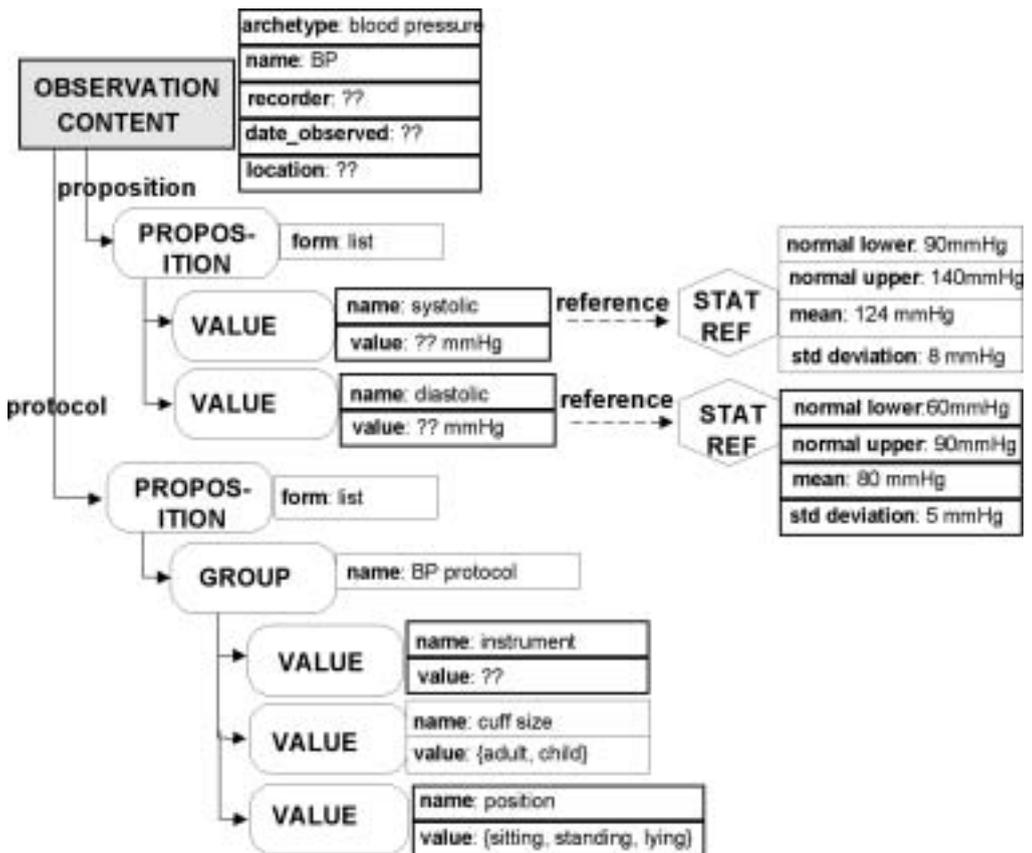

Figure 6: Blood Pressure Archetype including richer constraints

sentences such as "good English has went". Archetypes provide a flexible template for forming sensible sentences.

The main advantage of archetypes is that the clinical concepts (represented by the archetypes) are separated from the record management concepts (represented by the reference model). Thus, peak clinical bodies, such as colleges of physicians and government departments, can work on standardizing clinical content without worrying about the mechanics of record management. At the same time, it allows a single piece of EHR software to be built, which can handle numerous clinical concepts as well as being immune to changes in clinical knowledge.

2.4.4 Other EHR Standards Work

The first EHR architecture standard was CEN ENV12665, which was published in 1995 and was heavily based on the Good European Health Record project. This was replaced in 1999 by the CEN ENV13606 four-part EHR standard which still had its roots in the original GEHR, but added significant contributions from post-GEHR EU projects such as Synapses and EHCR-SupA, which addressed the important question of legacy systems and attempted to rectify some of the deficiencies of the earlier standard. The 13606 pre-standard had an improved architecture but still had significant limitations in terms of software implementability. This led to the decision in November 2001 to revise 13606 and to make it a full de jure standard upon completion of the revision, which is scheduled for January 2004. A major feature of the revision is the adoption of the GEHR archetype methodology. Work is also being done to merge the CEN EHR and openEHR reference models (the latter being a merger of the Australian GEHR and UK SynEx models).

3. IMPLEMENTATION

3.1 GEHR Trials

In order to test GEHR's ability to act as a common model for multiple health record systems, two trials funded by the General Practice Computing Group (GPCG) investigated the process of exporting clinical information from existing systems into a GEHR-based repository. These trials were run under the "Trials of IM/IT clinical integration activities within the health sector" RFT (Request for Tender). The two projects are described below.

- *OACIS project* (formally titled "Hospital to GP communication between non-GEHR and GEHR-compliant systems"): OACIS (Open Architecture Clinical Information System) is a commercial clinical information system adopted by the South Australian Department of Human Services to provide access to clinical data from South Australian hospitals. The GPCG project involved the transfer of approximately 15,500 de-identified microbiology, biochemistry, haematology and radiology test results from the South Australian OACIS system (DSTC and Flinders, 2001).
- *GP Software Integration project* (formally titled "Shared diabetes care in general practice"): This GPCG project looked at the transfer of medication lists, therapeutic precautions, problem lists and events for diabetes patients from Medical Director and Locum, two popular GP applications in Australia (Flinders and DSTC, 2001).

Automated translation from existing systems to a common EHR model, which these trials explore, is important to the success of a "shared EHR" network since a large amount of clinical data is stored in existing systems and the manual re-keying of this data to a "shared EHR" is not practical.

The software tools that were built for these two trials are explained in this section, and the lessons learnt are discussed in the next section.

3.2 Architecture

The overall architecture of the trial systems developed is shown in Figure 7. The primary storage unit in the system is the EHR node, which collects clinical information from a variety of sources. The data is stored in the EHR node as GEHR records in XML. The first trial (the OACIS project) looked at the scenario depicted in the bubble at the top-left corner. The pathology test results from the OACIS system were converted to the GEHR format, before being stored in the EHR node. The second trial (the GP Software Integration project) examined the scenario depicted in the middle-left bubble. At the patient consultation, the GPs used their local, clinical application to manage patient information. When they wanted to upload the information to the EHR node, they used the Conversion Tool, which queried the database of their local clinical application, converted the data into GEHR format, and uploaded the data into the EHR node. The conversion process from source clinical system to GEHR format is discussed in Sub-Section 3.3.

Doctors and other health care providers can then use their web browser to view the collected data in the EHR node, via an Access Point. The Access Point essentially acts as a web portal, and it is discussed in more detail in Section 3.4. The Access Point relies on other support services, including:
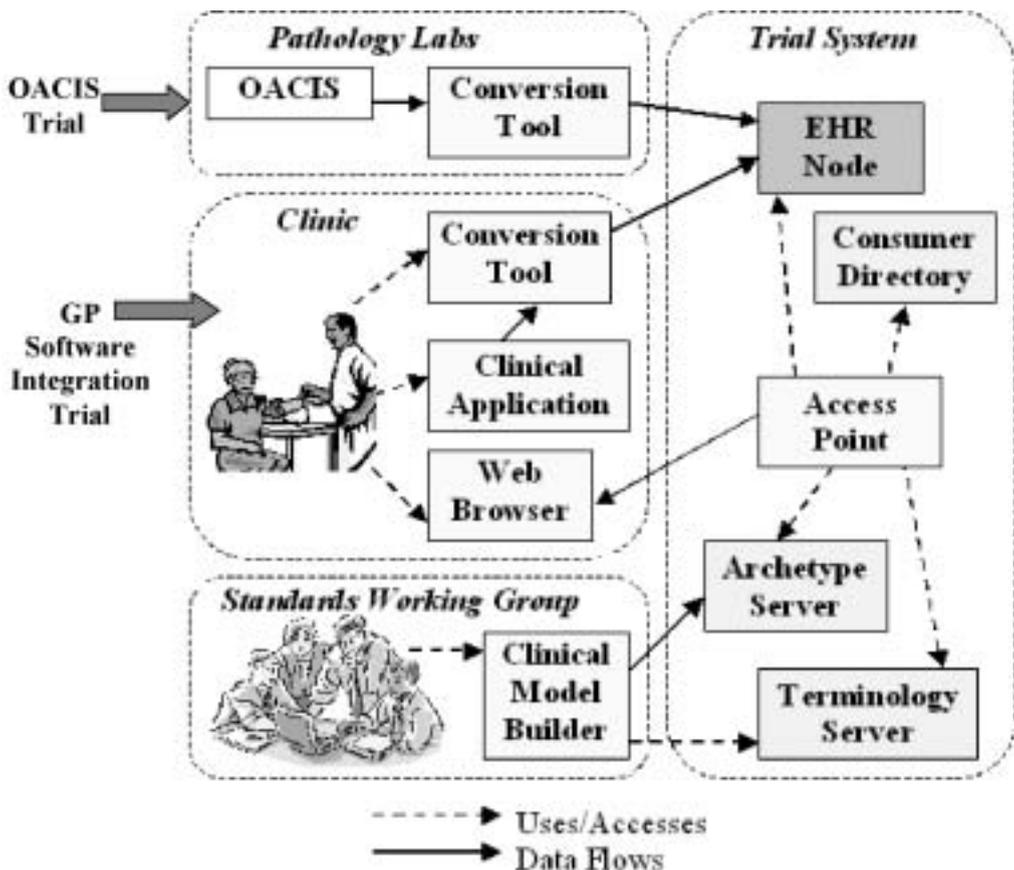

Figure 7: System Architecture of Trials

- A directory to look up the contact details and other information of health care consumers, or patients;
- Access to standard medical terminology sets; and
- A repository of archetype definitions.

As mentioned earlier, archetypes should be defined by standards working groups that are composed of people with clinical knowledge. A graphical user interface application called the Clinical Model Builder makes it easier to author archetypes and upload them to an Archetype Server for use in the system. This tool is discussed in Section 3.5.

3.3 Conversion Process

Although the OACIS project and the GP Software Integration project had different source systems and employed a different set of technologies (these differences are discussed shortly), they followed the same conversion process – namely:

1. Raw data is extracted directly from the source clinical systems.
2. This data is pre-processed to generate an XML format that corresponds directly to the extracted data.
3. In parallel, standard archetypes are designed for the type of clinical data being processed.
4. Given the XML data from the source clinical system (produced in Step 2), and standard clinical archetypes into which this data should be transformed (produced in Step 3), a mapping process is performed to define the relationships between the fields in each.
5. From the mapping process performed in Step 4, an XSLT (XML Stylesheet Language – Transformations) script is written, which is able to automatically transform the generic XML data into GEHR-compliant data conforming to the specified clinical archetypes.
6. The resulting XML-formatted GEHR data is then imported into the EHR repository.

The two trials used different sets of technologies to match the requirements of the scenarios they targeted. The OACIS project used Perl (Practical Extraction and Reporting Language) to parse the text data from the OACIS system to the intermediate XML format. Perl has good built-in capabilities for string manipulation, which proved useful in parsing the proprietary format of the OACIS data. Unix shell scripts were written to integrate the entire process in a single command. This approach was deemed adequate because it was envisioned that, in the OACIS scenario, a system administrator would be the individual to invoke the conversion process on large batches of data.

In the GP Software Integration project, the scenario involved a GP uploading the data into the EHR node, which required a more user-friendly process. A graphical user interface application was built, which stepped through the conversion process using a wizard. This tool was more sophisticated than that built for the OACIS project. For example, the user can specify the set of records that should be uploaded based on patient names, an optional time period and the types of data to be uploaded. A graphical view of the resulting GEHR-formatted XML data is provided, but the user can also view the underlying XML data, if necessary.

The internal design of the conversion tool (from the GP Software Integration project) used in this trial is shown in Figure 8. As shown in this diagram, the conversion tool consists of an Application Graphical User Interface (GUI) that calls upon various service classes to perform the tasks of extracting relevant clinical data, formatting the data, converting the data to GEHR-compliant XML and uploading the data to an EHR server.
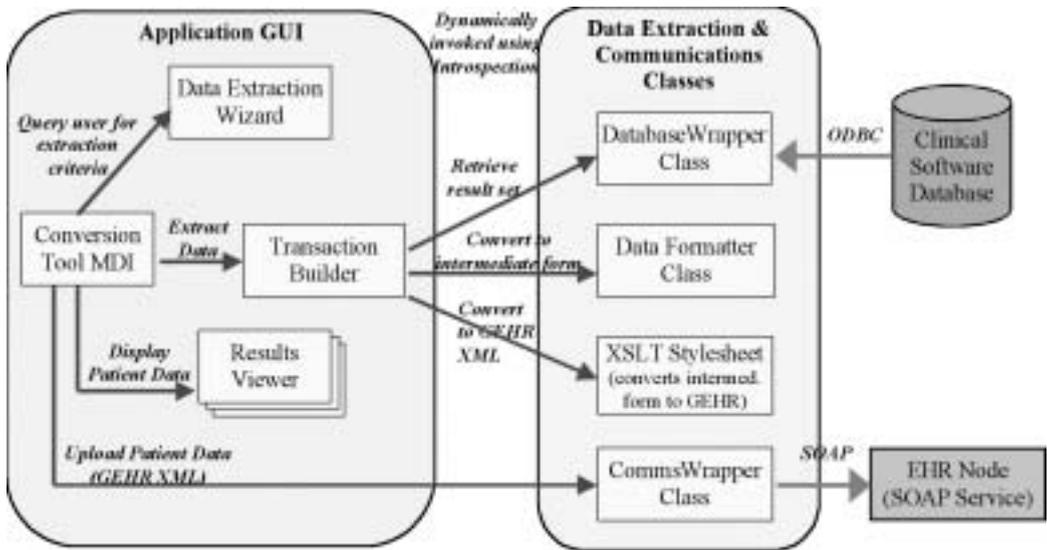
Figure 8: Internal Design of GEHR Conversion Tool

3.4 Web Portal

Once the EHR data from the different source clinical systems has been uploaded to the EHR node, it can be viewed through the web interface of the Access Point. This web portal provides a number of value-added services. For example, health care providers can search for patients in the system based on patient name, Medicare number, date of birth, and other fields. Figure 9 shows a page displaying the collated data on a single patient.

After selecting the patient from the results page, the user can view information about the patient's interactions with health care providers. For structured quantitative data, like the biochemistry test results from the OACIS system, the web portal can aggregate it to generate graphs, as shown in Figure 10. This feature makes it easier to see the trends in a patient's health.

3.5 Building Archetypes

Archetypes are formal constraint definitions that can be used to automatically validate EHR content. At the same time, it is intended that people with medical knowledge, who understand what makes clinical content valid, author them. With this in mind, a graphical tool called the 'Clinical Model Builder' was built to make it easier for clinicians to author archetypes. Figure 11 shows how a blood pressure archetype may look, when viewed with the Clinical Model Builder. While archetype authors using the clinical model builder do need to understand the principles of modelling and defining constraints, they do not need to understand or write the more complex XML syntax used by formal archetype definitions.

4. FINDINGS

The two trials discussed in this paper, namely the OACIS project and the GP Software Integration project, explored using GEHR as a common EHR model for data extracted from clinical systems. In performing these trials, we learnt a number of valuable lessons about importing data from existing clinical systems into a common EHR model. These lessons will be discussed in this section.
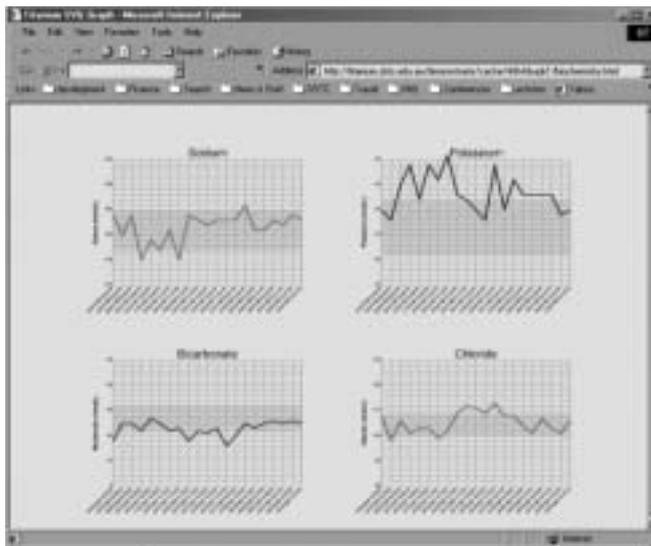
Figure 9: Patient Search Results Page

Figure 10: Graphs of Biochemistry Test Result

Figure 11: The Clinical Model Builder

4.1 Archetypes Approach

The trials showed that the archetypes approach does work. The GEHR reference model was able to represent data from two types of sources, namely GP practices and pathology labs, which have very different underlying data models. At the same time, a wide variety of clinical structures – from simple blood pressure values to highly structured biochemistry results – can be described using archetypes.

4.1.1 Designing Archetypes

A lesson learnt in designing archetypes was that archetypes needed to be generic and reusable, as well as clinically meaningful. If archetypes were designed to directly correspond to the data model of the clinical source systems, it would be pushing the data interoperability issues to the archetype level, rather than solving them. A single clinical concept needed to correspond to a single archetype, so it should be generic enough to handle the different contexts in which it can be used. For example, the archetypes that we designed for the pathology result can be reused for the different categories of pathology data (Bird *et al*, 2002).

4.1.2 Archetypes and XML

Another important lesson that we learnt in this trial is that an approach should not be driven by technology such as XML. It is better to first define the model and then figure out how to make XML work for you rather than defining the model first in XML and trying to figure out how to implement a specific requirement. We learnt this in relation to implementing archetype validation. Since the EHR content was in XML format, we thought that archetype constraints for the EHR content could be expressed using one of the numerous schema languages for XML and checked using existing tools. This proved to be wrong.

The first XML schema language considered was W3C's (World Wide Web Consortium) XML Schema (Thompson *et al*, 2001). The original idea was that the 'classes' in the reference model would be defined as complex types in XML Schema, and archetypes, which constrain the reference model, would map to 'class restrictions' in XML Schema. However, there are strict rules in using the restriction feature in XML Schema, which made it impossible to implement archetype constraints. For example, XML Schema's restriction feature does not allow you to apply a different set of constraints on separate items of a list of elements (e.g. the first item's name must be 'systolic' and the second item's name must be 'diastolic'). The ability to restrict lists in this way is fundamental to the way that archetype constraints work.

RELAX NG (Clark and Makoto, 2001), another XML schema language, was also investigated, but there were also difficulties defining archetype constraints in this language. Each archetype checks a fragment of an EHR XML document, which can be found anywhere in the document. The process of making RELAX NG schemas check constraints from locations that are not the start of the document proved to be too complex. Combined with the issue of there only being a small number of RELAX NG parser implementations available, this approach was abandoned.

Another language that was considered is Schematron (Jelliffe, 2002), a schema language that can search for patterns, such as error conditions, in an XML document and report them. Basically, Schematron schemas get translated to XSLT (Extensible Stylesheet Language – Transformations) scripts, consisting of 'If' statements whose conditions are expressed in XPath. This solution worked for simple archetypes like blood pressures, e.g. the first item's name is "systolic" and the second item's name is "diastolic", where the fields were mandatory. It proved impossible to express in XPath when fields or groups of fields in archetypes were optional or could occur multiple times, like challenge actions, e.g. fasting, in biochemistry tests. To overcome this problem, we considered writing XSLT scripts directly to validate archetype constraints. While this approached worked, the XSLT scripts performed poorly and were not easy to maintain, since they used recursion to overcome the lack of programmatic variables and loops in XSLT.

As a result of these issues, the XML schema languages were ultimately abandoned for the purposes of archetype validation. Instead, XML-encoded EHR data and archetype definitions will be read into the EHR system, which will validate the EHR data against the archetype constraints programmatically.

4.2 Difficulties in the Conversion Process

In both the OACIS and GP Software Integration projects, most data could be translated successfully. In the OACIS project, out of the total 15545 de-identified results, only three results were rejected because of data errors, and 146 were ignored because they were out of scope for the project. The GP Software Integration project was conducted differently. Clinicians typed in data for 18 fictitious patient contacts into the GP software from fictional scenarios, and this data was extracted from the databases of the GP software. All 18 patient contacts were extracted successfully.

Although most data were successfully translated, some data could not be translated well. The issues encountered in the conversion process are discussed next.

4.2.1 Conversion Issues

There were difficulties in the conversion process in both the OACIS and GP Software Integration projects because of the difference in approaches taken by the source clinical systems and GEHR. For example, we found the following issues:

- There were mandatory fields in GEHR that were not provided by the source systems. For example, for measured quantities, GEHR requires their units, e.g. cm, kg, and scientific

properties, e.g. length, mass, to be provided. Units were often provided except for some places, in which case, they had to be inferred from the context where the quantities were recorded. Scientific properties were not recorded at all, but they could be inferred from the units. Also, some contextual information about the data, such as the person recording the data, was not provided by the source systems, so these fields were marked as "unknown".

- The data in the source systems was stored in a less structured form than that of the archetypes. For example, in one GP software, information about a patient's allergies, family history and warnings were placed in free-form text fields. There was not enough structure to parse the fields sensibly to extract individual pieces of data. For example, only the name of the allergy and the associated reaction could be extracted from the allergy data, while the rest of data, including important information, such as when the allergy first came about, had to be placed in Comment fields in the archetypes. These pieces of data, therefore, cannot be queried or used in decision support systems. In addition, some mandatory fields in archetypes could not be filled, even though the actual data was present in the source systems.

- There were also differences in the way that data was grouped in the source and target format. For example, in the OACIS project's microbiology archetype, antibiotics are grouped by organism name, but the original OACIS data was grouped by each antibiotic name. While this issue was easily resolvable, it added to the complexity of the conversion process.

There were also a number of issues related to the quality of data in the clinical sources systems, such as:

- Some data was stored in places that were not considered intuitive. For example, one of the GP software stored family history, tobacco and alcohol related information in the Allergy table. In the biochemistry data of the OACIS system, if a patient was required to fast before a test, then the keyword "Fasting" is appended to the Units fields of the raw data in the test result (e.g. "mmol/L Fasting"). These problems became a source of some confusion in the mapping process.

- Some data that was mandatory in the data model of the source systems, was not always provided. For example, in the OACIS data, even though a 'master-panel' name was required to identify the type of test performed, there were a few rows, which did not include this. Because the type of test performed determined the structure of the data generated, these results could not be processed. Only three rows out of 15545 rows were thrown away, so it did not pose a serious problem.

- Some data was not placed in the appropriate data fields specified by the source systems, but in the less structured comment fields. In quite a number of rows in the OACIS data, the data fields, in which test results should have been placed, were in fact empty, while the test results were placed in the comment field. The content of these test results also had to be mapped to a Comment field in the GEHR archetype, meaning that this data will not be able to be easily queried or used in decision support systems. This problem occurred in 4115 rows, which is roughly a quarter of the OACIS data.

4.2.2 Discussion

The two trials give some valuable insight to the practical realities of introducing an integrated EHR network. The value of an integrated EHR network also depends on the quality of data provided by feeder systems as much as the common EHR model adopted. If the data from source systems lacks a standard structure, it will remain unstructured when it is imported into the "shared EHR". Thus, although a common EHR model like GEHR may have rich functionality and provide for sophisticated mechanisms like decision support systems and querying, these features cannot be

exploited if the source systems do not provide the data in the appropriate form. However, it should be noted that the source feeder systems examined in the trials were built based on different goals to that of GEHR. The existing systems were targeting more specific requirements for their operational context, so they did not record all information specified in GEHR, which addresses a broader set of requirements. The lack of structure in data does not pose a problem if the purpose of the data is simply meant for display, which seems to be the case for the source systems, whereas GEHR takes into account mechanisms like decision support systems and querying, which require more richly structured data. Once there is a business case for feeder systems to provide better data quality and structure, such as through a growing impetus for an integrated EHR network and increased use of decision-support systems, the issues discussed in the previous section may be resolved.

4.3 Challenges with a Two-level Modelling Approach

The two level "archetype" approach underlying GEHR is based on the premise that the reference model will remain fairly stable, while "archetypes" will be created and modified to enable the system to evolve over time with current medical practices and knowledge. Within the trials, we found that the archetype approach was flexible enough to capture the clinical information required. However, over the period of the two trials, the GEHR reference model went through a number of changes. To reflect the changes in the reference model, modifications to the core software components (such as the Clinical Model Builder and the EHR Instance Viewer) became necessary.

The changes in the reference model can mainly be attributed to the fact that GEHR as a model is still maturing, and consensus about the reference model is still forming. Although, in the long term, the "archetypes" approach may be more cost-effective, as software is shielded from changes in the clinical knowledge, a large initial investment is required to ensure that the reference model is correct, viable and sufficiently agreed upon. Furthermore, any changes to the reference model need to be carefully managed and versioned as the reference model provides the fundamental blocks of interoperability.

5. CONCLUSION

In this paper, we have described the latest Australian GEHR research and presented the outcomes of two General Practice Computer Group funded trials. We did this by first defining Electronic Health Records, comparing a number of approaches to building EHR systems and describing the archetype approach taken by the Australian GEHR research. We then described the implementation and findings of two Australian GEHR trials, which investigated the process of exporting clinical information from existing clinical systems into a GEHR-based repository.

The importing of large clinical datasets into a GEHR-based Health record, such as was done in the two trials discussed in this paper, is the first evidence we have that the GEHR approach to transforming EHR data can deal with real data of a reasonably diverse nature. This paper concludes, however, that it is very important that the issues uncovered and lessons learnt during these trials are fed into future work in this area.

5.1 Future Work

With the goal of applying the findings from the GEHR work to date, and heading towards harmonisation or convergence with other Health standards, such as CEN and HL7, CDA, the *open*EHR foundation, is developing a revised EHR model, referred to as the *open*EHR model. This model, which is largely based on the Australian GEHR research described in this paper, will form the underpinning of a trial, to be run by the Queensland Health department, under the banner of the

HealthConnect research program. Health*Connect* is the name of a concept to develop a national health information network, within Australia. This Queensland trial, among other things, will determine the suitability of the *open*EHR approach to be used as the EHR architecture for Health*Connect*.

ACKNOWLEDGEMENTS

REFERENCES

ALSCHULER, L., DOLIN, T., BOYER, S., BEEBE, C. (eds.) (2000): *HL7 Clinical Document Architecture Framework, Release 1.0*. ANSI-Approved HL7 Standard.

BIRD, L., GOODCHILD, A., HEARD. S. (2002): Importing clinical data into electronic health records: Lessons learnt from the first Australian GEHR trials. In: *HIC, Sydney, 2002*.

BHASALE, A., MILLER, G., REID, S., BRITT, H. (1998): Analysing potential harm in Australian general practice: an incident-monitoring study. *Med J Aust* 169:73–76.

BRANGER, P., VAN'T HOOFT, A., VAN DER WOUDEN, J., MOORMAN P., VAN BEMMEL, J. (1999): Shared care for diabetes: supporting communication between primary and secondary care. *Int J Med Inf* 53:133–42.

CLARK, J., MAKATO, M. (2001): *RELAX NG Specification*. OASIS Committee Specification. http://www.oasis-open.org/committees/relax-ng/spec-20011203.html Accessed 29 April 2003.

DOLIN, T., ALSCHULER, L., BEEBE, C., BIRON, P., BOYER, S., ESSIN, D., KIMBER, E., LINCOLN, T., MATTISON, J. (2001): The HL7 Clinical Document Architecture. *J Am Med Inform Assoc* 8:552–569.

DSTC, FLINDERS (2001): The OACIS-GEHR Transformation Process. *GPCG funded IM/IT Projects 2001* http://www.gpcg.org/publications/docs/projects2001/GPCG_Project2_01.pdf Accessed 29 April 2003.

FLINDERS, DSTC (2001): Exchanging diabetic patient data from GP software using GEHR. *GPCG funded IM/IT Projects 2001*.

GEHR (2000): *The GEHR Story* http://www.gehr.org/introduction/gehr_story.html Accessed 29 April 2003.

ISO (2002): *Health Informatics – Requirements for an Electronic Health Record Reference Architecture*. ISO/CD TS 18308.

JELLIFFE, R. (2002): The Schematron Assertion Language 1.5. http://www.ascc.net/xml/schematron/ Accessed 29 April 2003.

NATIONAL ELECTRONIC HEALTH RECORDS TASKFORCE (NEHRT) (2000):*A Health Information Network for Australia* http://www.healthconnect.gov.au/publications/publications.html#other_publications Accessed 29 April 2003.

KOHN, L., CORRIGAN, J. (eds.) (2000): *To Err Is Human: Building a Safer Health System*. Washington D.C: National Academies Press.

RECTOR A., NOWLAN, W., KAY S. (1992): Foundations for an electronic medical record. In: BEMMEL, J., McCRAY, A. (eds.) IMIA – *Yearbook of Medical Informatics 1992*. Stuttgart: Schattauer.

THOMPSON, H., BEECH, D., MALONEY, M., MENDELSOHN, N. (eds) (2001): *XML Schema Part 1: Structures*. W3C Recommendation. http://www.w3.org/TR/xmlschema-1 Accessed 29 April 2003.

BIOGRAPHICAL NOTES

Dr Linda Bird is a Senior Research Scientist at the CRC for Enterprise Distributed Systems Technology (DSTC), where she is one of the team leaders for the Titanium project. In this role, she conducts research and supervises prototype development in the area of Electronic Health Record inter-operability, using XML and web-based technologies. Dr Bird has been involved in a number of jobs and consultancies requiring her data modelling, data design and meta-modelling skills – including the Australian Government's HealthConnect-openEHR project and data modelling consultancy to the Brisbane North Division of General Practice's Coordinated Care Trial ('TeamCare'). Dr Bird has a Bachelor of Information Technology with first class honours and a University Medal from the University of Queensland. She obtained her Ph.D. in 1997 for her research in "Data Reverse Engineering".

Linda Bird

Dr Andrew Goodchild is a Senior Research Scientist at the DSTC, and works in a wide variety of roles, including: project management, research or software architecture. Andrew's main skills are on projects with a need for secure and scalable enterprise data access and exchange. Andrew holds a Ph.D. in Computer Science from the University of Queensland. In recent consultancies he has consulted to large enterprises with high volume needs such as the US Patents and Trademarks Office, the Australian Defence Department and the Australian federal health records initiative: HealthConnect. Andrew has knowledge in technologies such as very large databases, distributed systems, J2EE, object-oriented design, web services, XML and security.

Andrrew Goodchild

Zar Zar Tun is a Research Scientist at the Distributed Systems Technology Centre (DSTC), based in Brisbane, Australia. She performs research and development for the Titanium project, which investigates how to provide secure, lightweight, enterprise-wide data access, particularly focusing on electronic health records.

Zar Zar Tun