Netw Model Anal Health Inform Bioinforma (2013) 2:213–223
DOI 10.1007/s13721-013-0038-6

ORIGINAL ARTICLE

An ontology-driven approach to mobile data collection applications for the healthcare industry

Gabriela Henriques · Laura Lamanna · Daniel Kotowski · Hlomani Hlomani · Deborah Stacey · Philip Baker · Sherilee Harper

Received: 15 November 2012 / Revised: 28 June 2013 / Accepted: 29 July 2013 / Published online: 20 August 2013

Abstract Technology has often been associated with improvements in many domains. This is particularly true in the medical and healthcare industry. This is a field where data collection is performed on a daily basis. With the advent of mobile technology, several methodologies for data collection have been adopted to reduce the cost and time expended on data collection. The focus of this paper is a proposed ontology-based framework that has the ability to build a shared repository of surveys that can be used for data collection. The paper discusses *iCollect*, a first instantiation of the framework in the form of a survey application built for the Indigenous Health Adaptation to Climate Change (IHACC) project.

G. Henriques (✉) · L. Lamanna · D. Kotowski · H. Hlomani · D. Stacey
School of Computer Science, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada
e-mail: ghenriqu@uoguelph.ca

L. Lamanna
e-mail: llamanna@uoguelph.ca

D. Kotowski
e-mail: dkotowsk@uoguelph.ca

H. Hlomani
e-mail: hhlomani@uoguelph.ca

D. Stacey
e-mail: dastacey@uoguelph.ca

P. Baker
Public Health Agency of Canada, 600 Southgate Drive, Guelph, ON N1G 4P6, Canada
e-mail: bakerp@uoguelph.ca

S. Harper
Department of Population Medicine, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada
e-mail: harpers@uoguelph.ca

Keywords Ontology · Software engineering · Healthcare · Data transfer · Data collection · Mobility · Data security

1 Introduction

The medical and healthcare industry is characterized by the constant need for data collection and analysis. In fact, in this industry it is a standard practice for large amounts of data to be collected on a daily basis such as in emergency rooms, medical clinics and in the field for biomedical and public health research purposes. The immediate challenges with regards to data management (including data collection and analysis) include the cost (temporal and monetary) involved in the process and the sensitivity of the data that also impacts on such tasks as storage, access, and transfer. Traditional data collection methods, mostly conducted with pen and paper, do little to help address the aforementioned challenges.

Technological investments and various data collection methods have been implemented in recent years. Most of these can be attributed to the apparent "boom" of mobile technology. While these mobile devices come to the aid of the industry, they are also characterized by some limitations including, but not limited to, small screen size, limited memory capacity and battery life. These limitations drive the development of better and more efficient software solutions.

The Guelph Ontology Team (GOT) has conducted research in knowledge engineering (KE) (particularly in ontologies) and software engineering (SE) with a focus on flexibility, reusability and efficiency (Hlomani and Stacey 2009). In this paper, we discuss *iCollect*, an application developed for survey-based data collection in healthcare. The current implementation of the system is simple; it alleviates the need for paper-based surveys. However, we

aim to expand it to use knowledge engineering techniques, so that not only will it allow for dynamic survey creation, but it will also be useful to many fields outside of healthcare. The healthcare industry is a great place to start as there is a growing need for semantic description of information being collected and used within the field as seen in (Xiang et al. 2012), (Zillner and Sonntag 2012) and (Li et al. 2012). We examine an approach that will allow for flexibility, extendibility, and a generic method for the gathering of health-related data that helps address various gaps associated with the existing methods. We envision that, in the future, this approach can be easily expanded into other domains with similar data collection requirements. We propose a framework that has the ability to manage a shared repository of questions and surveys that can be used for the data collection.

2 Background

2.1 Ontologies

Formally, *ontologies* are a conceptualization of a domain of interest (Gruber 1993). In simpler terms, they are structures which formally represent knowledge. These structures are defined using description logics, which lead to the ability to actively reason on these structures. When reasoned upon, the ontology may discover new relationships as well as allow for the querying of existing relationships (Guarino et al. 2009). Ontologies are comprised of `Classes` which represent a concept or a physical entity in the domain of interest. The link or relation between any two classes is established through `Properties`. The property which links the two classes denotes that these classes are a member of the same set. Often ontologies are confused with taxonomies. They are similar to taxonomies in their ability to represent structural hierarchy. However, an ontology's knowledge representation is not only limited to the notion of subsumption. In addition, ontologies are distinct in their ability to place restrictions on classes (to define conditional class membership) and their ability to be reasoned with (Jepsen 2009).

A trivial but easily understandable example is described in (Jepsen 2009). In the example (depicted in Fig. 1), we are able to represent a hierarchical depiction of all food types including their food instances in both ontology view and hierarchy view. However, with ontologies, we are able to leverage their power and pool together food types with a common property(ies) through the definition of a restricted class (either anonymous or named). To give an example, we may be able to identify fruits, vegetables and cheeses that are orange in colour.

2.2 Software engineering and ontologies

Concepts such as modularization, distribution, reuse, and integration of software components define the concerns of software engineering. The greater the need to both extend and automate these tasks, the greater the relevance of the use of ontologies as a conceptual model for these tasks and other software engineering components (Hesse 2005). From the computer science perspective, ontologies have their roots in the knowledge engineering domain (Happel and Seedorf 2006). Recent trends, however, show an

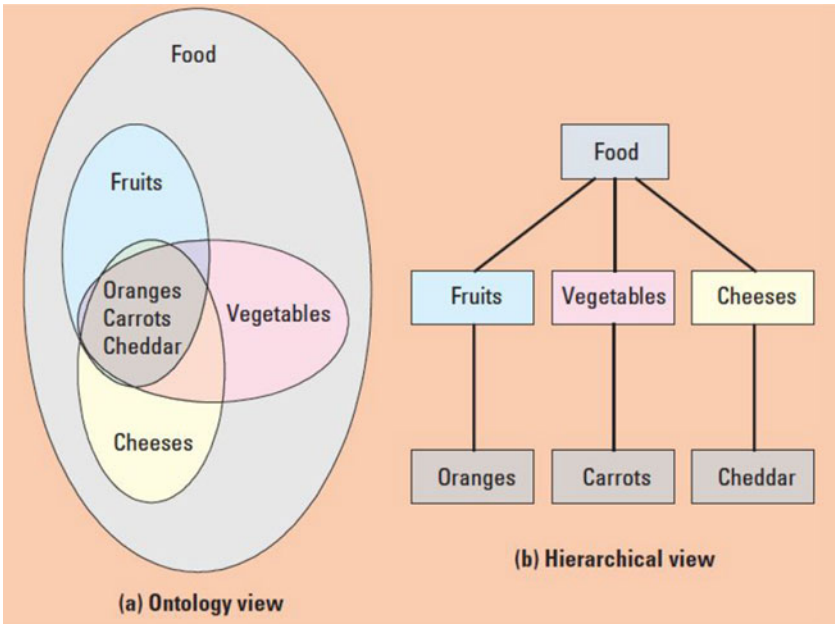

Fig. 1 Distinction between an ontology and a taxonomy. Source: Guarino et al. (2009)

overlap between SE and KE as evidenced by the recent adoption of ontologies as either software engineering artifacts or as part of the development strategy. We see in recent software development methodologies (e.g. model driven development and ontology-driven software development or engineering) a bias towards models as the main knowledge base from which inferences can be made (Hesse 2005). An ontology is, in fact, a conceptual model. There is also evidence of ontologies being used in other intricate processes of the software development cycle such as requirements elicitation (Kaiya and Saeki 2006). There are various ontology-based applications that have been developed for the healthcare industry: a syndromic surveillance application to detect disease outbreaks (Buckeridge et al. 2008), and extracting meta-data from images to reason and detect early stages of lymphoma in patients (Zillner and Sonntag 2012). These applications are based on the data previously collected in clinics and hospitals.

Proper software engineering is very relevant especially with regards to applications targeted for mobile devices. Ontologies are relevant in this discussion since they are the models that we will use to provide flexibility and extensibility in our framework.

2.3 Data collection and data transfer

In the past decade, there have been many advances in the medical and healthcare industry and with advances there comes a rise in patient‘s expectations (Pope and Mays 1995). Data collection and transfer still remain an important aspect in the industry; these tasks are required daily in hospitals and medical clinics and are necessary procedures for data analysis and research. Many data collection methods have been developed for the different fields within the healthcare industry in order to improve data retrieval (Pope and Mays 1995; Declich and Carter 1994). However, the format in which the data is being collected and then later transferred can impact the final result of the task at hand (Brown 2001). Data collection can be time-consuming and costly, especially in circumstances that require gathering large amounts of data for analysis.

The ubiquitous nature of data collection in the healthcare industry demands a fast, efficient and low-cost data collection methodology. In the healthcare industry there are various types of data being collected and various data collection techniques. One of these methods is known as self-reporting; this usually involves face-to-face communication between the participant and a healthcare professional (or whoever is gathering the data). A common example of this method can be seen when a patient goes for a yearly checkup to their doctor's office. Usually the doctor has a set of questions to ask the patient and if the patient is feeling sick then they report their symptoms to the doctor. Some of this information is currently stored on a computer, however, there is still a great deal of paperwork involved when visiting the doctor's office. As well, face-to-face communication is not always effective if the doctor is busy typing in the patient's answers to a computer with their back towards their patient.

Data is not always collected in only one area. There are some forms of data collection that involve gathering information in various locations. Having the ability to collect the data and have it transferred to one location that is then accessible to all the locations involved is ideal.

Many forms of data collection still involve pen and paper due to its mobility. This can be seen especially during research. This can be time-consuming and inefficient; sometimes the data collected is not entered into the system until the end of the experiment (Brown 2001). One disadvantage of this method is that the data can only be analyzed at the very end (Brown 2001). Having the ability to promptly analyze the data being collected can provide the benefit of determining whether the researcher is satisfied with the trend in the data and whether any changes with the collection need to be made during the research. Having this knowledge available only at the end can be costly.

The healthcare industry is composed of many subareas that contain different requirements in terms of how their data are retrieved and analyzed. However, they all share the notion of the need for effective data collection and transfer formats. Often, the term data transfer can mean inputting data from paper to another device such as a computer. If a form existed where the data can be collected on a device and automatically transferred onto a computer, the data could be saved in various formats. Data transfer can then include moving data from a device to a computer, server, data storage centre, or even to the Cloud. It can also encompass transferring the information directly to the desired format such as XML or a spreadsheet thus meeting the requirements of the task at hand.

Although the current preferred format of data collection is electronic, there are still limitations in data exchange, file transfer formats and the security of transmitting confidential data (Hariri et al. 2012).

2.4 Current technologies

With the rise of mobile technologies and growing access to the Internet, new ways of collecting health information have emerged. There have been several proposed systems that offer methods for digitizing and collecting health data and information as seen in the works of Kilkarni and Argawal (2008), Rahbar (2010), and Aanensen et al. (2009).

A variety of systems that use mobile technology have been designed for health and emergency alerts, tracking medical supplies, recording injuries of people affected during the Thailand and Cambodia disasters, asthma attacks and inhaler usage and infectious diseases (Freifeld et al. 2010). It is clear that there is a large potential for these technologies to positively impact the data collection process.

Another example of a mobile system proposed by Aanensen et al. (2009) allows multiple ground workers to use mobile devices to collect environmental data such as soil pH, sample temperature, whether the soil contains bacterium (*Xenorhabdus nematophila*), and GPS locations. Data can then be wirelessly uploaded to a website where it is processed and analyzed.

A similar technology to that used by Aanensen et al. (2009) was proposed by Kulkarni and Argawal (2008), as well as Rahbar (2010). But Rahbar (2010) goes further by not only collecting data but has a system that makes suggestions about the level of healthcare needed for end users based on the severity and frequency of their symptoms. Kulkarni and Argawal propose this kind of system for third world countries and use it as a recommendation system to help field-diagnosed patients. The system described by Rahbar is meant to advise a patient on where to receive proper healthcare. Both these systems are used to collect patient information, analyze it, and then present some sort of recommendation. However, due to its implementation and infrastructure, there is a lack of flexibility and data interoperability. Data rigidity makes it difficult to use data sets from these separate systems together and thus prevents these systems from achieving more of an impact.

We can see from existing systems that the limitation is not the technology; it is the method in which the data are being stored and processed. The major limitation of these systems is the specificity of their use; many of the operations are not flexible and for this reason it would be difficult to reuse such applications for similar tasks.

The approach examined in this paper allows for flexibility, extendibility, and a generic method for the gathering of health data, and helps address various gaps associated with the existing methods.

3 Proposal

Data collection and transfer can be costly and inefficient if not done properly. We propose a system that will provide for an efficient and effective method for collecting data in various sectors in the healthcare industry. An ontology-driven application will allow for an effective and efficient data collection and transfer system. The idea for this system came from one of the projects developed by GOT: *iCollect*.

3.1 The *iCollect* project

iCollect is a survey application built for the Indigenous Health Adaptation to Climate Change (IHACC) project for the Public Health Agency of Canada. IHACC (2012) has brought a multinational, interdisciplinary team together to develop an understanding of the vulnerability of remote Indigenous health systems to climate change. The program reflects needs identified by partners in the community, government, and Indigenous organizations during pilot research with Inuit (Canada), Batwa Pygmy (Uganda), and Shipibo and Shawi populations (Peru).

Each region experiences site-specific research, training and intervention activities while contributing unique data, forming the basis of pilot adaptation interventions and adaptation planning. The project's data collection concentrates on food security, water security, and vector-borne disease in a changing climate. Within these foci, attention is directed to the differentiation in vulnerability among children and elders, the importance of Indigenous knowledge, and the role of globalization and resource development.

This research addresses a significant deficit in understanding the health dimensions of climate change among Indigenous populations and, to our knowledge, is the first program to place explicit emphasis on implementing and monitoring adaptation interventions. The validated approach offers best practice guidance for other initiatives, creates community and scientific adaptation leaders with expertise in Indigenous health and climate change, and demonstrates the importance of Indigenous knowledge for adaptation, empowering communities to manage the health effects of climate change.

iCollect was developed as a data collection tool to ensure efficiency of survey administration and the streamlining of data into a centralized database accessible by all international teams. The process for survey administration would usually involve a paper survey which would be filled out in each remote region, shipped to the regional 'home university' and the data entered by hand—a lengthy process. The *iCollect* tool digitizes this process via an iPad survey application and enables teams to upload computerized data via an internet connection, saving time, money and keypad fatigue. A systematic and standardized system of data collection minimizes the likelihood of human error, increasing the quality and utility of the data. A digital collection tool also reduces the training required for the survey administrators, enabling less experienced administrators to conduct a survey with fewer mistakes and expanding the survey's capacity.

The application is currently supported on iPad devices, however, it contains the modularity necessary to support

other tablet devices such as those based on *Android*. The purpose of the application is to simplify everyday tasks that researchers in the field will be performing, allowing them a portable method for collecting data so that they may have the ability to carry out their tasks in various locations. The application also provides a rapid form of data transfer, so that once all the survey information has been gathered for one day, the answers can be directly stored in their data storage site. This has the benefit of securely storing confidential data for future analysis.

The current *iCollect* application is the first version of the system. It is currently not ontology based, but has the potential to be so. Putting an ontology behind the current *iCollect* system will allow this survey system to be utilized by a variety of different projects. An ontology will add the ability to classify surveys and survey questions, giving flexibility to the application by allowing for dynamic creation of new surveys or using a predefined survey set based on input or requirements provided by the user. This set of requirements will be used as parameters in the reasoning aspect of the ontology; the ontology will reason which survey or set of questions to display to the user based on the requirements that were given.

3.2 Objective

Using the notion of a survey-based method of data collection, a system can have a repository of surveys and questions. This repository can contain various types of surveys that are used for different purposes in the healthcare sectors; it does not necessarily have to include only one type of questionnaire set. As well, by having a repository, it is possible to have a uniquely designed survey to meet the requirements specified by the user. It will also allow surveys to be created on the fly based on user input and reasoning facilitated by the ontology.

Our goal is to develop an ontology-based application that will take some input from the user and determine how to design a survey that is best suited for the scenario (set of requirements) presented. The ontology is an important part of the application, as it is responsible for classifying the type of questions or surveys that are currently stored in the repository. The ontology will interact with a reasoner that will determine which set of questions are most appropriate based on some parameter set that has been provided. Ontologies are based on Description Logic, a knowledge representation formalism designed primarily for describing and reasoning about structural knowledge. The ability of ontologies to describe structural knowledge about the nature of questions and surveys will allow a conceptual model of data collection to be developed that can serve many different applications within the healthcare domain and will allow communication between this domain and other domains (e.g. economic, social) that may have data needed by a healthcare application.

Having the application run through a tablet device can allow for portability and ease of use for the user. A tablet is typically the size of a piece of paper; it is small enough that it can be carried around everywhere and big enough that the user has clear vision of the task at hand. Smartphones are also a viable option that can be used for the data collection, however, the tablet screen size allows for a more visually appealing system. The screen size of a tablet is similar to the size of a notebook allowing for ease of use and visibility.

Through collecting all survey answers on the mobile application, data transfer is quick and easy; there are many options that can be used in how or where the data is transferred. For example, data can be transferred into a spreadsheet (convenient for many public health sectors, as the data can then be used for analysis). Data can also be transferred to other formats (XML, PDF), or locations (servers, repositories).

To illustrate the objective of our proposed system, a simple scenario has been provided. The scenario represents an example of one of the ways the application could be used to collect data in healthcare and is meant to provide a high-level overview of the benefits that the ontology brings if the application were to be used as described below.

3.3 Scenario

3.3.1 Setting: a doctor's office in a medical clinic

A doctor receives a middle-aged male patient with type 1 diabetes for a regular yearly checkup at his clinic. The patient was required to furnish his basic information at the front desk before starting his appointment. The doctor is using a tablet to record necessary information he must collect from the patient. The doctor pulls up the patient's information through the tablet and his basic profile is presented. A basic profile can include such information as gender, age, health card number, and known health issues. The application can then allow the doctor to either proceed to commence the appointment, view historical information on previous visits or update the patient's profile information. The doctor is ready to start the appointment so he selects the start option. The application pulls up necessary questions that are typically required for a middle-aged male's regular checkup. As well, since the patient's profile has indicated that he has type 1 diabetes, sections are included that are indicated for those patients with type 1 diabetes. This survey set is individualized for that particular patient and takes into account other factors along with basic profile information that was mentioned above.

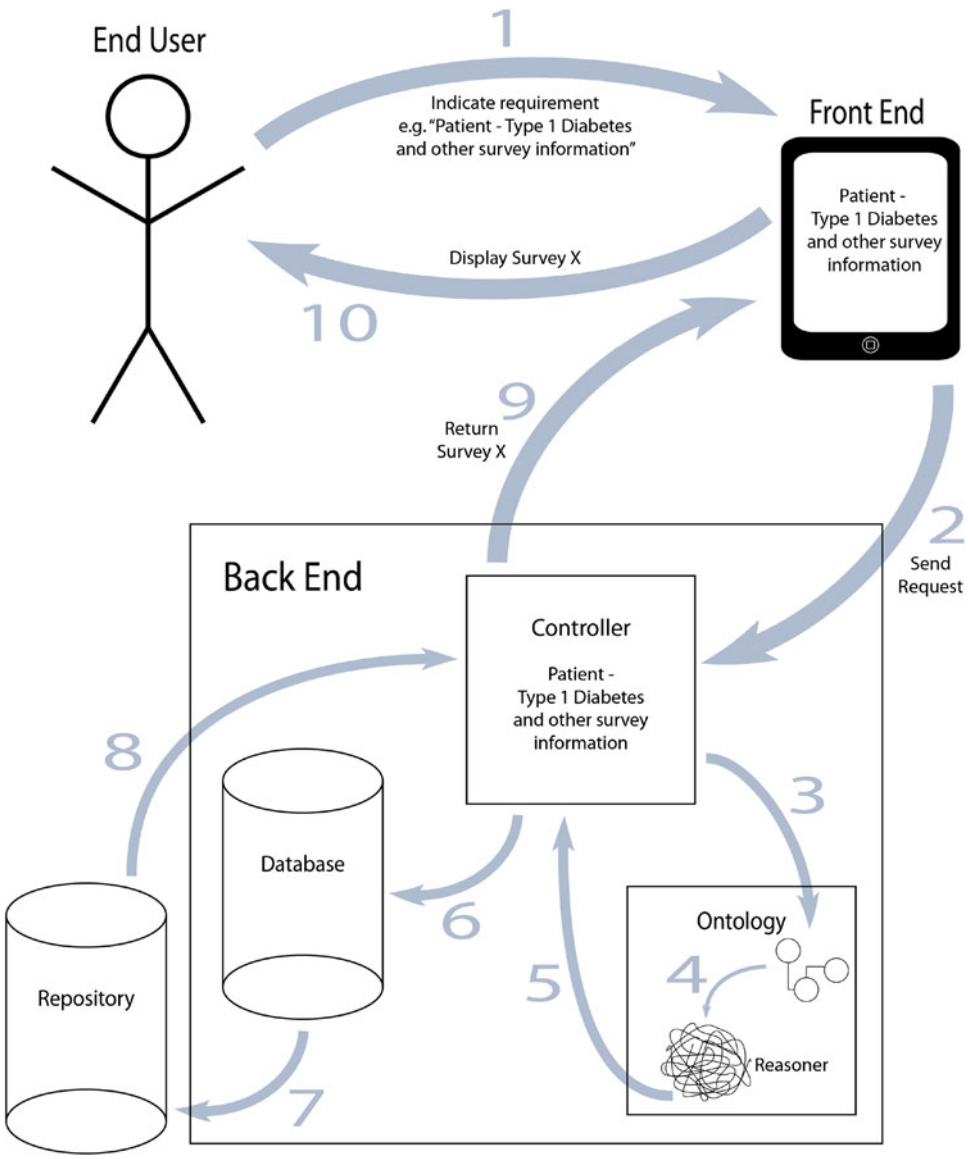

Fig. 2 System architecture

Some other factors may include demographic location, geospatial information, etc. The survey set is designed to meet the needs of the doctor–patient appointment. No two patients will necessarily have the same set of questions since the application will design the appointment survey set based on the parameters entered about each individual in the current appointment. The question sets will also change over time as new information about various conditions (such as diabetes) is available.

4 Methodology

4.1 Architecture

Figure 2 displays the system architecture for the survey-fetching procedure in the proposed ontology-driven data collection system. This further illustrates the scenario described in the previous section. The following steps detail the approach and are each automated by the various system components.

4.1.1 The user indicates the requirements

The end user interacts with the front end (the interface of the program). This is the client application designed to work with the centralized survey retrieval system. Depending on the situation, the end user could be a doctor, secretary, lab technician, or any other professional requiring dynamic survey content in the field. The interface's design could be different based on the user's occupation. For example, a clinical secretary may use a desktop application whereas a doctor may use a tablet. The user inputs the data for the patient or subject.

4.1.2 The front end sends the request to the back end

The "front end" application sends the data to the centralized "back end" system. This is a server or set of servers configured to accept and process requests from the clients. The data are sent directly to the "controller" part of the "back end" which is responsible for parsing through the request data and taking appropriate actions.

4.1.3 The controller sends the data to the ontology-based subsystem

After the user input is processed by the controller, it is passed to the ontology-based subsystem. In conjunction with Step 4, the knowledge base defined by the ontology is used to classify the data.

4.1.4 The reasoner classifies the data

Based on the ontological classification derived in Step 3, the reasoner decides on the requirements for the desired survey which is output back to the controller. An existing reasoner such as FaCT++, Hermit or Pallet would be utilized.

4.1.5 The reasoned response is sent back to the controller

The controller receives the requirements for the needed survey from the ontologies and reasoner. These specifications are used to formulate a database lookup for the needed survey set.

4.1.6 The requirements for the survey content are used in a database search

The controller then sends the parameters received by the ontology-based subsystem to the database as a query. This can be a local or remote database, depending on the circumstances. This step can potentially consist of several queries to different databases holding more specialized information (for example a query for a patient with diabetes may redirect to a database maintained by a diabetes association). These databases can be constantly updated with the most recent research on a subject or domain so that appropriate questions and references to the medical literature can be made available to practitioners.

4.1.7 The required survey or survey parts are fetched from a remote database/repository

The results of the database query are then used to fetch the appropriate surveys/questions from other databases or repositories. As mentioned in Step 6, depending on the size of the system, the database lookup may require that the surveys are retrieved from several different sources.

4.1.8 The survey data are returned to the controller

The survey data are sent back to the controller. Here, it is processed and translated into a form readable by the requesting client system.

4.1.9 The full survey is returned to the client

The complete survey is sent back to the requesting client. The front end parses the received survey data and uses it to populate the user interface.

4.1.10 The survey is displayed to the user

The survey questions are then displayed to the end user. The client system receives their input and collects and records the data.

4.2 Ontology survey meta-model

In this section, we focus on the core element of the system (the ontology) which serves as its backbone. The ontology, hereby referred to as the *survey ontology* is the meta-model upon which survey descriptions are based. This survey meta-model serves three purposes:

1. *Survey validation* describes the criteria that needs to be satisfied for any survey to be considered valid and consistent.
2. *Question type description* provides descriptions of the structure and semantics of any given question.
3. *Data context* this gives semantic properties to data to allow for a deeper understanding of its context.

These features are realized during various stages of the survey's life cycle. For example, survey validation typically occurs during the survey creation process, while the data context is defined during the survey creation but persists throughout the lifecycle of the survey. To illustrate these features, consider a UML depiction (Fig. 3) of some of the classes[1] of the ontology. In as far as validation is concerned, a survey can be considered to be valid as long as it is within the confines of the structure prescribed in the `Survey` class. Taking advantage of ontologies, one could then constrain the definition of a survey through, for example, existential restrictions (`someValueFrom` in OWL or `some`). These restrictions (existential restrictions) indicate the existence of at least one relationship along a

[1] Not to be confused with the implementation classes for the tool, but only pertains to the ontology design.

given property (in this case the `hasQuestion` property) to a specific class (e.g. `Question`). This can also be read to the effect that, for a survey to be valid, there should be at least one question definition or *hasQuestion some Question*. This is formally denoted as $\exists$ *hasQuestion Question*. Existential restrictions do not, however, mandate that the only restrictions that can exist alongside the property must be relationships to a given class (e.g. the filler class for `hasQuestion` must be `Question` class). This means that there is no guarantee that the questions would conform to the `Question` structural constraints. The restriction can therefore, be extended to be universal restrictions (`allValuesFrom` in OWL or `only`). This would be denoted as $\forall$ *hasQuestion Question*.

For simplicity, and for the purposes of our discussion, the `Location` class, particularly its subclass (country), has been constrained to an enumeration of predefined countries. This could, for example, be countries where the surveys can be taken. This definition can, however, be extended to inherit definitions from other already existing geographic or geospatial ontologies through the importation of such ontologies. Likewise the `QuestionType` class is an enumeration of an exhaustive list of question types [e.g. *multiple choice single* (MCS) *select* type of question—where the user is offered multiple options but only allowed a single valid answer choice, *multiple choice multiple* (MCM) *select*—where the user is offered many options and can select as many options as required, and last but not least *single choice* (SC)]. In Sect. 4.1 we mentioned that user parameters could be used for information retrieval; a practical example of that would be where a query for surveys is constrained to surveys taken in a specific country.

The guiding principle of the ontology design was to have a structure that is general enough to cover a wide range of surveys and question types while also allowing for flexibility in customization for individual specific circumstances. This extendibility would allow the ontology to be reused to help adapt it to appropriate domains.

5 Discussion

A mobile/tablet application utilizing ontologies and distributed computing can provide many benefits for healthcare professionals. Many current methods for health data collection still involve using more traditional forms of data collection; a doctor/nurse pushing a cart with a computer around to each patient to record or update information on how the patient is doing, or pen and paper based surveys

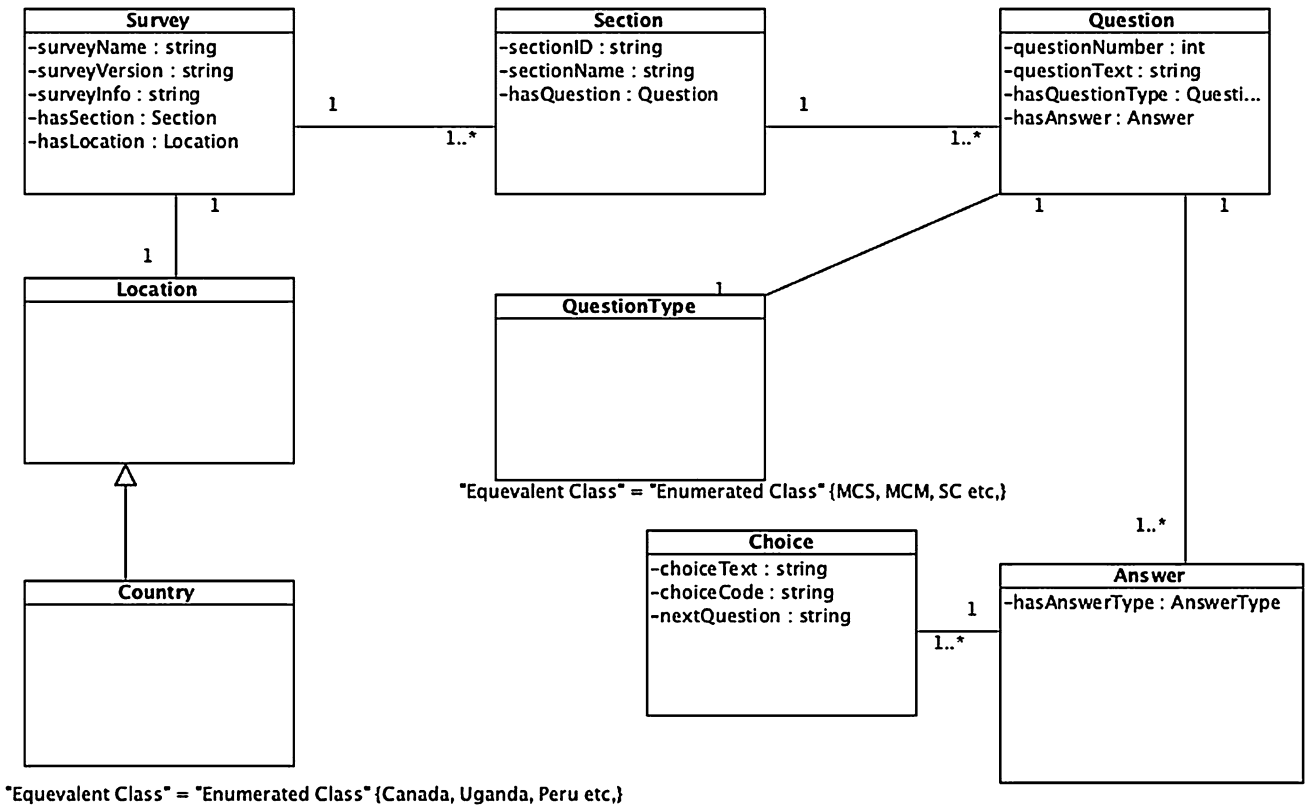

Fig. 3 UML representation showing some of the classes of the survey ontology. While cardinality aspects of the relationship between the classes are shown, other axioms (and complete listing of the ontology) are not given

collecting health information from communities. Even laptops can be intrusive if they distract the doctor from the patient. Mobile devices linked to servers that supply dynamic expert information with regards to data collection (the providence behind the data and the need to gather particular data) can move health information into a new paradigm characterized by flexibility, adaptability, portability and transparency.

Ease of data collection will empower many aspects of health research. Timely, appropriate data will move into the system for analysis and healthcare professionals will be able to capture, store and share (when appropriate) data in a cost-effective and efficient manner.

The below sections will discuss how the proposed application addresses some important concepts in the data collection process in terms of mobility and data security.

5.1 Mobility

In sociology, the term mobility is defined as the movement of people in a population, as from place to place, from job to job, or from one social class or level to another. By adapting this term to data collection, we refer to mobility as being the movement of data from place to place; from a person to an application to a database, or from server to server. One of the goals of this proposal is to aid users by easing the process of data collection and providing a smooth transition of data from one place to another; in other word, we want to provide the user with the mobility they need.

This is a standalone application; it does not depend on integration with other applications or services to be fully functional. As previously mentioned, the main objective of the application is to aid a user in collecting survey-based data. Healthcare is a large field, with a wide range of different data types and data sources; it is difficult to determine one universal data collection method. Healthcare is a good example of a domain with the need for survey-based data collection, which is commonly seen wherever there is a need for collecting data from a third party. For example; gathering data from a patient to a physician or nurse, gathering data from a group of individuals to a researcher (as done in the IHACC project). When it comes to data being collected from a third party (most commonly face-to-face interaction), the proposed framework can meet most of these needs.

There are two streams of mobility worth discussing: the flexibility provided through the ontology and the portability provided using a standalone mobile application.

Depending on the users needs, there are two different ways the application can be utilized. The application can come with a set of pre-built questions and surveys, or the questions and surveys can be attained from another source of data such as a server, the cloud, etc. In the first way, the ontology would be locally grabbing the appropriate set of questions or surveys that should be displayed to the end-user based on the requirements they have entered. This method does not rely on internet connectivity, and is the recommended form of using the application when performing field research in underdeveloped or rural communities.

There is a need for a better method of data collection in underdeveloped as well as rural communities, where there is little to no internet connectivity. If we take a look at how data was originally collected for the IHACC project, surveyors were going door-to-door gathering data on pen and paper. Another example of research done using the same method can be seen in "Accelerating Health Service and Data Capturing Through Community Health Workers in Rural Ethiopia" (Damtew 2012), where their main form of data capture is done through pen and paper surveys. Using the application with a pre-loaded set of surveys and questions its use does not become limited due to the lack of internet connectivity and is convenient to use in rural or underdeveloped communities where technology is not up to date. This form of mobility where a bundle of paper is replaced by a tablet or smart phone is better for the user (data collector), and could serve to intrigue participants or patients as well. Although you become limited with the amount of survey data that can be stored on the tablet device, you still attain the advantage of having a survey built based on your needs.

The other method would require more set-up with the application and would rely on internet connectivity to allow the ontology to attain the appropriate questions and surveys stored on a third-party location. The advantage of this is that it would allow a larger amount of storage for questions and surveys, and has the potential for opening collaboration opportunities between other parties to form a shared repository of survey data. As well, by giving the user this option, this demonstrates the flexibility of the application in that the user is not restricted to using a set of uploaded surveys repeatedly, and can benefit from obtaining the most appropriate survey to meet their needs at any time.

Once the survey or question set has been selected, the questions will be displayed and the user can commence their data collection; once all data has been collected it is saved in the appropriate format on the collection device. How the data is extracted and saved is based on the discretion of the user; storage and security of data are discussed in the section below.

5.2 Data storage security

5.2.1 Current implementation

In the current implementation of the *iCollect* application, after a questionnaire has been completed, the answers are

saved as XML codes in a file on the iPad. The file can be moved from the iPad by connecting to a computer and uploading them via iTunes. The files can also be moved to a central database via secure FTP.

This process is cumbersome, especially to those who are less technologically inclined. However, it is beneficial in a few ways. Firstly, it ensures complete confidentiality of the data. The software developers or distributors are not involved beyond installing and updating the software. The users are in complete control of the data, and so they take all responsibility for it. This ensures no legal issues involving other parties that need to access the data. A second benefit to this method is that the users do not need to rely on a third party to design or host their database. They have complete control over what is done with the raw data after it is uploaded from the iPad. Thirdly, this process is beneficial when out in remote locations. The application does not rely on having an internet connection to upload the data, so the files (collected data) can remain locally on the iPad until paired with a computer to upload the files.

This process also comes with some consequences, which can be addressed in the proposed framework. Currently, because the application stores the data unencrypted, there is always the possibility that if the iPad is stolen or compromised, the survey answers will be at risk. But one of the largest drawbacks, as mentioned above, is that this method is not good for less tech-savvy users. The users may not be comfortable with taking the files off of the iPad and uploading them to a database themselves. They may need a third party to design and host their database if they do not have the expertise. In addition, if the users make a mistake while taking the files off of the iPad or transferring them, they could potentially lose the data. This could be solved by passing the responsibility of the data onto a third party, which will be discussed in the section below.

The current version of the system was designed to collect data for a simple set of surveys, so that it could replace the inefficient pen-and-paper method. Our proposed application, incorporating ontologies, would address these issues.

5.2.2 Proposed framework

For the proposed framework, several features could be improved so that the above drawbacks are addressed. To address data security, the survey results could be encrypted while being stored on the mobile device. This ensures that the data will not be at risk in case the device is compromised. To avoid putting the responsibility of uploading the data files on the users, the application could instead automatically upload the survey results when the device is connected to the internet. This would help avoid accidental loss of data and be more helpful for those that are not as technologically-inclined. This would require that the mobile devices have access to a secure internet access point occasionally. It would also eliminate the need to pair a computer with the mobile device. The application could automatically upload the results to a database set up for the users. This database could use cloud storage either as a permanent storage location or as a temporary waypoint for the data. This would put the responsibility for data security and confidentiality on a third party, the cloud storage provider. Cloud storage, however, introduces new issues such as service charges and the question of where the data are being stored.

5.3 Future directions

The current *iCollect* application is being adapted to support an ontology-based server-side subsystem to enable the application to dynamically create new surveys based on user requirements. Domains outside of the original IHACC one are being investigated.

Acknowledgments *iCollect* development was funded by the IHACC project. The software is open source.

References

Aanensen DM, Huntley DM, Feil EJ, al Own F, Spratt BG (2009) EpiCollect: linking smartphones to web applications for epidemiology. Ecol Commun Data Collect 4(9):e6968

Brown SJ (2001) Research data collection and analysis. US Patent 6,196,970, 6 Mar 2001

Buckeridge DL et al (2008) Understanding detection performance in public health surveillance: modeling aberrancy-detection algorithms. J Am Med Inform Assoc 15(6): 760–769

Damtew Z (2012) Accelerating health service and data capturing through community health workers in rural ethiopia. Knowl Eng Ontol Dev

Declich S, Carter A (1994) Public health surveillance: historical origins, methods and evaluation. Bull World Health Organ 72(2):285–304

Freifeld CC, Chunara R, Mekaru SR, Chan EH, Kass-Hout T, Ayala Iacucci A, Brownstein JS (2010) Participatory epidemiology: use of mobile phones for community-based health reporting. PLoS Med 7(12):e1000,376

Gruber TR (1993) Toward principles for the design of ontologies used for knowledge sharing. In: International journal of human-computer studies. Kluwer Academic Publishers, Dordrecht, pp 907–928

Guarino N, Oberle D, Staab S (2009) What is an ontology? In: Staab S, Rudi Studer D (eds) Handbook on ontologies, international handbooks on information systems. Springer, Berlin

Happel H, Seedorf S (2006) Applications of ontologies in software engineering. In: Proceedings of international workshop on semantic web enabled software engineering (SWESE06), pp 1–14

Hariri S et al (2012) The HPV Vaccine Impact Monitoring Project (HPV-IMPACT): assessing early evidence of vaccination impact on HPV—associated cervical cancer precursor lesions. Cancer Cause Control 23(2):281–288

Hesse W (2005) Ontologies in the software engineering process. In: Proceedings of the workshop on enterprise application integration (EAI 2005)

Hlomani H, Stacey DA (2009) An ontology driven approach to software systems composition. In: International Conference of Knowledge Engineering and Ontology Development, pp 254-260

IHACC: Indigenous Health Adaptation to Climate Change (IHAC) (2012). http://www.ihacc.ca

Jepsen TC (2009) Just what is an ontology, anyway?. IT Prof 11:22–27

Kaiya H, Saeki M (2006) Using domain ontology as domain knowledge for requirements elicitation. In: 14th IEEE international requirements engineering conference RE06, vol 167342, pp 189–198

Kulkarni S, Agrawal P (2008) Smartphone driven healthcare system for rural communities in developing countries. In: Proceedings of the 2nd international workshop on systems and networking support for health care and assisted living environments, HealthNet '08. ACM, New York, pp 8:1–8:3

Li J, Guo L, Handly N, Mai A, Thompson D (2012) Semantic-enhanced models to support timely admission prediction at emergency departments. Netw Model Anal Health Inf Bioinforma 1:161–172

Pope C, Mays N (1995) Reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. BMJ 311(6996):42–45

Rahbar A (2010) An e-ambulatory healthcare system using mobile network. In: Seventh international conference on information technology: new generations (ITNG), pp 1269–1273

Xiang Y, Fuhry D, Kaya K, Jin R, Catalyurek U, Huang K (2012) Merging network patterns: a general framework to summarize biomedical network data. Netw Model Anal Health Inf Bioinforma 1:103–116

Zillner S, Sonntag D (2012) Image metadata reasoning for improved clinical decision support. Netw Model Anal Health Inf Bioinforma 1:37–46