Evaluating Health Information Systems Using Ontologies

Shahryar Eivazzadeh

Abstract

Background Several frameworks try to address the challenges of evaluation of health care information systems, by offering models, methods, and guidelines about what to evaluate, how to evaluate, and how to report the evaluation results. The evaluation aspects, i.e. what to evaluate, suggested by model-based evaluation frameworks are usually recommended for universal application and do not include case-specific aspects. In contrast, for the evaluation aspects extracted by frameworks that rely on elicitation of requirements, while they include case-specific aspects but they are limited to evaluation aspects foreseen by case users in the early phases of system development. At the same time, the elicitation approaches extract different evaluation aspects from each case, make it challenging to compare or aggregate the evaluation of a set of heterogeneous health care information systems together.

Objectives It is needed to find a method for harvesting and unifying different aspects (or whats) of evaluation of health care information cases in a heterogeneous environment. The method needs to both aggregate the evaluation aspects extracted by elicitation of user's requirements and accommodate recommended evaluation aspects by a model based evaluation framework.

Method Based on available literature and methods in semantic networks and ontologies, a method (called Unified Evaluation by Ontology Integration (UVON)) was developed that harvests and aggregates the evaluation aspects of several health care information systems, from both elicitation and model-based sources, into a tree-style ontology structure. Then the ontology structure was used to extract aggregated evaluation aspects that are originated both from requirements elicitation and health technology evaluation models.

Results The method was applied in Future Internet Social and Technological Alignment Research (FI-STAR), a project of seven cloud-based e-health applications being developed and deployed across European Union countries. That resulted in an ontology structure which reflects the internally declared required quality attributes in the FI-STAR project, which then was extended to accommodate the quality attributes recommended by Model for ASsessment of Telemedicine applications (MAST) evaluation framework. The evaluation aspects were extracted from the output ontology and were used for the evaluation of all FI-STAR cases, both as separate cases and as a whole.

Conclusions The suggested method can provide a way for organizing, unifying, and aggregating the evaluation aspects of heterogeneous health care information system implementations that are suggested by model-based and elicitation-based approaches. The method maintains to be systematic, context-sensitive, and less subjective across a heterogeneous set of health care information systems.

Keywords

Health Care Information Systems; Ontologies; Evaluation; Technology Assessment, Biomedical

1 Introduction and Background

the war against evaluation failure

At least in one aspect, the evaluation of health care information systems matches well with their implementation: they both fail very often (Littlejohns et al., 2003; Kreps and Richardson, 2007; Greenhalgh and Russell, 2010). In consequence, in the absence of an evaluation that could deliver insight about the impacts, an implementation cannot gain accreditation to join the club of successful implementations. Beyond the reports in literature on the frequent accounts of this kind of failures (Greenhalgh and Russell, 2010), the reported gaps in the literature (Chaudhry et al., 2006), and the new emerging papers that introduce new ways of doing the health care information system evaluation (Yusof et al., 2008), including this paper, can be interpreted as a supporting indicator that the attrition war on the complexity and failure-proneness of health care information system is still ongoing (Yusof et al., 2008). The opposite front of the complexity and failure-proness of evaluation are models, methods, and frameworks that try to address what to evaluate, how to evaluate, or how to report the result of an evaluation. In this front, this paper tries to contribute to the answer of *what to evaluate*.

user-based success criteria

As a cornerstone for evaluation, there comes our interpretation of what things constitute success in health care information systems. A body of literature has been developed concerning the definition and criteria of being a successful health technology, where the criteria for success go beyond the functionalities of those systems (Holden and Karsh, 2010; Berg, 2001). Models similar to Technology Acceptance Model (TAM), being applied to health technology context, formulate this success as the end-users acceptance of a health technology system (Hu et al., 1999). Success of a system, hence acceptance of a health care information system, can be considered as the usage of that system when the using is voluntary, or it can be considered the overall user acceptance when the using is mandatory (Goodhue and

Thompson, 1995; Ammenwerth et al., 2006).

To map the definition of success of health care information systems into the real world cases, the evaluation frameworks have emerged (Ekeland et al., 2012; Yusof et al., 2008). These frameworks with their models, methods, taxonomies, and guidelines are supposed to capture parts of our knowledge about health care information systems. This knowledge enables us to evaluate those systems and it enlists and highlights the elements of the more effective, more efficient, or less prone to failure evaluation processes. Evaluation frameworks, more specifically in the summative approach, might address what to evaluate, when to evaluate, or how to evaluate (Yusof et al., 2008). These frameworks might also elaborate more by addressing evaluation design, the way to measure the evaluation aspects or how to compile, interpret, and report the results (Talmon et al., 2009).

Evaluation frameworks offer a wide range of components for designing, implementing, and reporting an evaluation, where suggestions or guidelines for finding out the answer to *what to evaluate* are among those components. The answer to *what to evaluate* can range from the impact on structural or procedural qualities to more direct outcomes such as the overall impact on patient care (Ammenwerth et al., 2004). For example, in the STARE-HI statement which provides guidelines for the components of a final evaluation report of health informatics, the 'outcome measures or evaluation criteria' denote the *what to evaluate* question (Talmon et al., 2009).

To identify the evaluation aspects, the evaluation frameworks can take two approaches, top-down or bottom-up. Frameworks that take the top-down approach try to specify the evaluation aspects through instantiating a model in the context of an evaluation case . Frameworks that focus on finding, selecting, and aggregating evaluation aspects through interacting with users, i.e. being the so-called user-centered, have taken the bottom-up approach.

top-down or bottom-up approach

In the model-based category, the Technology Acceptance Model (TAM) and Technology Acceptance Model 2 (TAM2) models, with their wide application in different disciplines including health care (Holden and Karsh, 2010), begin from a unique dimension of *behavioral intention to use (acceptance)* , as a determinant of success or failure, and go to expand it to *perceived usefulness* and *perceived ease of use* (Davis, 1989; Holden and Karsh, 2010), where these two later dimensions can become the basic constructs of the evaluation aspects. The Unified Theory of Acceptance and Use of Technology (UTAUT) framework introduces four others —still similar— determinants which are *performance expectancy*, *effort expectancy*, *social influence*, and *facilitating conditions*. Here, the first two can become basic elements for evaluation aspects, but the next two might need more adaptation to be considered as aspects of evaluation for a health care information system.

Some model-based frameworks extend more by taking into consideration the relations between the elements in the model. The Fit between Individuals, Task and Technology (FITT) model includes the *task* element beside the technology and individual elements. It then extends by creating a triangle of *fitting* relations between these three elements. In this triangle, each of the elements or the interaction between each two elements is a determinant to success or failure (Ammenwerth et al., 2006), and therefore each of those six can construct an aspect for evaluation . The Human, Organization, and Technology Fit (HOT-fit) model builds upon DeLone & McLean Information Systems Success Model (DeLone and McLean, 2004) and extends more by including the *organization* element beside the technology and human elements (Yusof et al., 2008). This model also creates a triangle of *fitting* relations between those three elements.

The outcome-based evaluation models , such as the Health IT Evaluation Toolkit provided by the Agency for Healthcare Research and Quality, consider very specific evaluation measures for evaluation. For example, in the above mentioned toolkit, measures are grouped as domains, such as *efficiency*, and there are suggestions or examples of possible measures for each domain, such as *percent of practices or patient units that have gone paperless* (Cusack and Poon, 2007).

In contrast to model-based approaches, bottom-up approaches are less detailed on the evaluation aspects landscape, instead, they form this landscape by what they elicit from stakeholders. Requirement engineering, as a practice in system engineering and software engineering disciplines, is supposed to capture and document, in a systematic way, the user's needs of a to-be-produced system (Cheng and Atlee, 2007). The requirements specified by requirement documents, as a reflection of user's needs, determine to a considerable extent what things are needed to be evaluated at the end of system deployment and usage phase, in a summative evaluation approach. Some requirement engineering strategies apply generic patterns and models to the extracted requirements (Cheng and Atlee, 2007), hence show some resemblance, in this regard, with the model-based methods.

As an advantage, the elicitation-based approaches, such as the requirement engineering, directly reflect the case-specific user's needs in terms of functionalities and qualities. They enumerate and detail the aspects that are needed to be evaluated, all from the user's perspective. Evaluation aspects that are specified through requirement engineering process can dynamically be added, removed, or changed due to more interaction with users or other stakeholders at any time. Their changes, such as getting more detailed or more generic, is due to new findings and insights, new priorities, or the limitations that rise in the implementation of the evaluation.

The advantages in the requirement engineering approach come with the cost of some limitations in compare to model-based methods. Most of the requirement elicitation activities are accomplished in the early stages of the system development when the users do not have a clear image of what they want or do not want in the final system (Friedman and Wyatt, 2006, page 64). However, a model-based approach goes beyond the requirements expressed by the user's of a specific case by presenting models that are summaries of the past experiences in a wide range of similar cases and studies.

the challenge of comparing heterogeneous things

BEING case-specific by using requirement engineering processes has a side effect and that is the different sets of evaluation aspects that it elicits from each case which can be even heterogeneous to each other. Model-based approaches might perform more uniform in this regard, as they try to enumerate and unify the possible evaluation aspects through their models imposing a kind of unification from the beginning. However, still there exists a group of studies asking for measures to reduce the heterogeneity of evaluation aspects in these approaches(Ekeland et al., 2012).

Heterogeneity makes evaluation of multiple cases or aggregation of individual evaluations a challenge. In a normative evaluation, comparability is the cornerstone of the evaluation (Bürkle et al., 2001), in that sense that things are supposed to be better or worse than each other or than a common benchmark, standard, norm, average, or mode, in some specific aspects. Without comparability, the evaluation subjects, at best, can only be compared to themselves in their different stages of life course (longitudinal study).

In health technology, the challenge of heterogeneity for comparing and evaluation can be more intense. The health technology assessment literature applies a very inclusive definition of *health technology*, which results in a heterogeneous evaluation landscape. The heterogeneity of evaluation aspects is not limited to the heterogeneity of actors and their responses in a health setting, but it also includes the heterogeneity of health information technology itself. For example, the glossary of health technology assessment by the International Network of Agencies for Health Technology Assessment (INAHTA) describes health technology as the 'pharmaceuticals, devices, procedures and organizational systems used in health care.' (Facey, Karen et al., 2006), which conveys how the intervention is packaged in chemicals, how it is supported by devices, how it is organized as procedures running over time, or how it is structured or supported by structures in the organizational systems. Similar inclusive and comprehensive definitions can be found in other studies (Kristensen, 2009; Draborg et al., 2005). This heterogeneous evaluation context can create problems for any evaluation framework that tries to stretch and be applicable for a diverse set of health technology implementations. This heterogeneity can cause challenge for an evaluation framework in comparing evaluation aspects (Busse et al., 2002) and in consequence in summing up reports (Lampe et al., 2009), the creation of unified evaluation guidelines, and even in the evaluation of the evaluation process.

heterogeneity of health technology in definition level

By extracting the lowest common denominators amongst evaluation subjects, hence, creating a uniform context for comparison and evaluation, we can tackle the challenge of heterogeneity in elicitation based evaluation approaches.

evaluation aspects as the lowest common denominators of the evaluation objects

Vice versa, the evaluation aspects in an evaluation framework suggest the common denominators between different elements. The lowest common denominator, as its mathematical concept suggests, expands to include the elements from all parties where the expansion has been kept to the lowest possible degree.

Usually, there is a trade-off and challenge between the universality of an evaluation aspect, i.e. how common it is, and its relativeness, i.e. how low and close to the original elements it lays. When the scopes are different, their non-overlapped areas might be considerable, making it a challenge for finding the common evaluation aspects. Further more, the same concepts might be perceived or presented differently by different stakeholders (Ammenwerth et al., 2003). Also, different approaches usually target different aspects to be evaluated, as a matter of focus or preference.

It is possible to merge the results of model-centered and elicitation-centered approaches. The merged output provides the advantages of the both approaches and cover some of the challenges and shortcomings in each approach by the other one.

solution: merging the approaches

THE aim of this paper is to address the question of *what to evaluate* in a health care information system by proposing a method (called Unified eValuation using ONtology) which constructs evaluation aspects by organizing quality attributes in ontological structures. The method deals with the challenges of model-based evaluation frameworks by eliciting case-specific evaluation aspects, adapting in evaluation aspects from some model-based evaluation frameworks, and accommodating new cases that show up with time. The method can address heterogeneity by unifying different quality attributes that are extracted from one or more evaluation cases. This unification can happen by an arbitrary degree of balance between similarities and differences, regarding the needs of the evaluation implementation. As a proof of applicability of the proposed method, it has been instantiated and used in a real world case for evaluating health care information system.

summary of contributions

In the rest of the paper, the research method that resulted in the UVON method is described in Section 2. The result, i.e. the UVON method, is covered in Section 3.1, while its application in the context project is covered in Section 3.2. The rationale behind the method is discussed in Section 4 and the possible extensions and limitations come in Section 4.1 and Section 4.2. The Section 5 would conclude the paper.

paper structure

2 Method and Materials

THE FI-STAR project is a pilot project in e-health systems funded by European Union (EU). The Evaluation of the FI-STAR project has been the major motive, the empirical basis, and the testbed for our proposed evaluation method, i.e. the UVON method (to be described in Section 3). The FI-STAR is a project within the Future Internet Public-Private Partnership Programme (FI-PPP) and relates to the Future Internet (FI) series of technology platforms. The project consists of seven different e-health cloud-based applications, being developed and deployed in seven pilots across Europe. Each of these applications serves a different community of patients and health professionals (FISTAR Consorium, 2014) and have different expected clinical outcomes. FI-STAR and its seven pilot projects raised the challenge of finding an evaluation mechanisms that can be utilized both to evaluate each project and as well to aggregate the result of those evaluations as the evaluation of the whole FI-STAR project.

the FI-STAR project

A general review of the existing evaluation frameworks was done. Existing model-based evaluation frameworks, which usually suggest universal quality attributes for evaluation, could not cover all the quality attributes (i.e. evaluation aspects) reflected by the requirement documents of the pilot projects in FI-STAR. Even if there was a good coverage of the demanded evaluation aspects, still there was no guarantee that they can maintain the same degree of that good coverage for the future expansions of the FI-STAR project. On the other hand, the requirement documents from the FI-STAR project were not supposed to be the ultimate sources for identifying those quality attributes. It was speculated that there could exist other relevant quality attributes that are captured in the related literature or embedded in other, mostly model-based, health care information system evaluation frameworks. Regarding the above mentioned reasons, it was decided to combine quality attributes both from the FI-STAR sources and a relevant external evaluation framework. In order to find other relevant evaluation aspects, a more specific review of the current literature was performed, where it was then more focused to find an evaluation framework of health care information systems that matches enough with the specifications of the FI-STAR project. The review considered MAST framework (Kidholm et al., 2012) as the candidate evaluation framework. This evaluation framework was supposed to cover the quality attributes that were not indicated in the FI-STAR requirement documents but were considered needed to be evaluated in similar projects. These extra quality attributes are suggested by opinion of experts and background studies (Kidholm et al., 2012). Nevertheless, the quality attributes extracted from this framework were needed to be integrated with the quality attributes extracted from the FI-STAR requirement documents.

Regarding the heterogeneity of FI-STAR's seven pilot projects, the evaluation mechanism needed to extract common qualities from different requirement declarations and unify them. A review of the related literature showed that the literature on ontologies refers to the same functionalities, i.e. capturing the concepts (quality attributes in our case) in a domain and their relation (Noy and McGuinness, 2001). It was considered that sub-class and super-class relations, and the way they get represented in ontology, can bring unification to the heterogeneous quality attributes that exist in our evaluation case. Regarding the possible future expansions of the FI-STAR project, this utilization of the ontological structures needed to be systematic and easily repeatable.

3 *Results*

A method was developed to organize and unify the captured quality attributes through requirement engineering into a tree-style ontology structure and to integrate that structure with the recommended evaluation aspects from another evaluation framework. The method was applied for the seven pilots of the FI-STAR project which resulted in a tree-style ontology of the quality attributes mentioned in the project requirement documents and the MAST evaluation framework. The top 10 nodes of the tree-style ontology were chosen as the 10 aspects of evaluation relevant to the FI-STAR project and its pilot cases.

the resulted ontology and the top 10 nodes

3.1 The UVON Method for Unifying the Evaluation Aspects

METHODICAL capture of a local ontology (Uschold, 2000) from the quality attributes, i.e. evaluation aspects ontology, and reaching unification through its tree structure nature is the primary strategy behind our method. Therefore, a method is introduced, which is called UVON to underline *Unified eValuation* of aspects as the target and *ONtology* construction or integration as the core algorithm. The ontology construction method presented in this paper is a simple, semi-automated method, configured and tested against FI-STAR project use cases. The UVON method does not try to introduce a new way of ontology construction; but it focuses on how to form and utilize a local ontology (Uschold, 2000; Choi et al., 2006) out of the quality attributes of a system, in the sake of finding out what to evaluate. In this regard, the ontology construction in the UVON method is a reorganization of common practices, as such as those introduced by Noy and McGuinness (2001).

case-specific considerations in the method

THE ontology structure, in its tree form, is the backbone of the UVON method. Modern ontology definition languages can show different type of relations, but for the sake of our method here, we only use the *is of type* relation. The *is of type* relation can also describe as pairs of parent and child, super-class and sub-class, or general and specific relations. This kind of relation creates a direct acyclic graph structure which is or can be converted to a tree form. In this tree, the terms and concept are nodes of that tree and branches consists of those nodes connected by *is of type* relations. The tree has a root, which is the super-class, parent, or the general form of all other nodes. Traditionally, this node has been called the *thing* (Noy and McGuinness, 2001), probably due to the fact that any thing can be categorized as a *thing*.

Figure 1 is an example about how this ontology structure can look like. All the nodes in this picture are quality attributes, except the leaf nodes at the bottom which are instances of health care information system. While going up to the top layers in the ontology, the quality attributes become more generic, at the same time aggregating and unifying their child nodes.

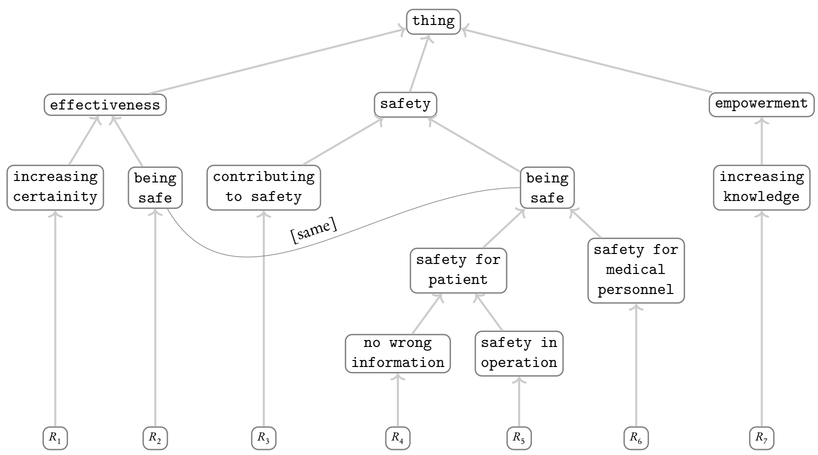

Figure 1: An example snapshot of the output ontology while running the UVON method

THE UVON method is composed of three phases α, β, and γ (Figure 2). In the first phase, all quality attributes elicited by requirement engineering process are collected in an unstructured set that is respectively called α set. In the next phase (β) and based on the α set , an ontology is being developed by the UVON method, which is called β(beta) ontology. In the next step, if the ontology gets extended by an external evaluation framework (as discussed in the method) then it is called γ (gamma) ontology.

construction of the β ontology

THE β ontology construction begins with a special initial node (i.e. quality attribute) that is called *thing*. All the collected quality attributes are going to begin a journey to find their position in the ontology structure, beginning from the *thing* node and going down the ontology structure to some certain points that is specified by the algorithm. This journey is actually a deep-first tree traversal algorithm (Tarjan, 1972) with some modifications. To avoid confusion in the course of this algorithm, a quality attribute that seeks to find it position is called the *traveling quality attributes* or Q_t.

The first quality attribute just needs to add itself as the child of the *thing* root node. For the rest of quality attributes, each checks to see if there exists any child to the *thing* node, where that is a super-class (superset, super-concept, general concept, more abstract form, etc.) to the traveling quality attribute (Q_t). If such a quality attribute exists (let's say Q_n) then the journey continues by taking the route through that quality attribute. The algorithm examines the children of Q_n (if exists any) to see if it is a sub-class to any of them (or they are super-class to Q_t).

The journey ends at some point because of the following situations: If there is no child for a new root quality attribute (Q_n) then the traveling quality attribute (Q_t) should be added as a child to this one and its journey ends. That is the same if there exist children to a new root quality attribute (Q_n) but none of them is a super-class and nor a sub-class to our traveling quality attribute. Beside these two situations, it is possible that no child is a super-class, but one or more of them are the sub-class of the traveling quality attribute (Q_t). In this situation, the traveling quality attribute (Q_t) itself becomes a child of that new root quality attribute and those child quality attributes move down become children of the traveling quality attribute (Q_t).

To keep the ontology as a tree, if a traveling quality attribute (Q_t) finds more than one super-class children to itself in a situation, then it should replicate (fork) itself into instances, as many as the number of those children, and go through each branch separately. It is important to note that logically this replication cannot happen over two disjoint (mutually exclusive) branches.

It is also possible to inject new quality attributes in between a parent node and children, of course if it does not break sub-class or super-class relations. This injection can help to create ontologies that the nodes in each level of the tree have a similar degree of generality and each branch of the tree grows from generic nodes to more specific ones.

This customized deep-first tree traversal algorithm, which actually constructs a tree-style ontology instead of just traversing that, is considered semi-automated as it relies on human decision in two cases. The first case is when it is needed to consider the super-class to sub-class relations between two quality attributes. The gradual development of the ontology through the UVON method spreads the decision about super-class to sub-class relations across the course of ontology construction. The unification between heterogeneous quality attributes (nodes) is the result of accumulating these distributed decisions which are embodied as super-class to sub-class relations. Each of these relations (i.e. decisions) makes at least two separate quality attributes closer together by representing them through more generic quality attributes.

In the second case of human decision or subjectivity, one can inject a new a new quality attribute to the ontology tree while that quality attribute is not explicitly mentioned in the requirement documents. This injection is only allowed when that quality attribute summarizes or equals a single or a few sibling quality attributes that are already in the ontology. The injection can improve clarity of the ontology. It can also help adjusting the branches of the ontology tree to grow to some certain height which can be helpful when a specific level of the tree is going to be considered as the base for creating a questionnaire. This adjustment of branch height might be needed If a branch is not tall enough to reach that specific level which then none of the quality attributes in that branch gets presented in the questionnaire. Also, if a quality attribute is very specific in compare with other quality attributes in that level of the tree then the questions in the questionnaire become inconsistent in their degree of generality. This inconsistency can be handled by injecting more generic quality attributes above the existing leaf node in the branch. All the above mentioned benefits come with the cost of subjectivity in introducing a new quality attribute.

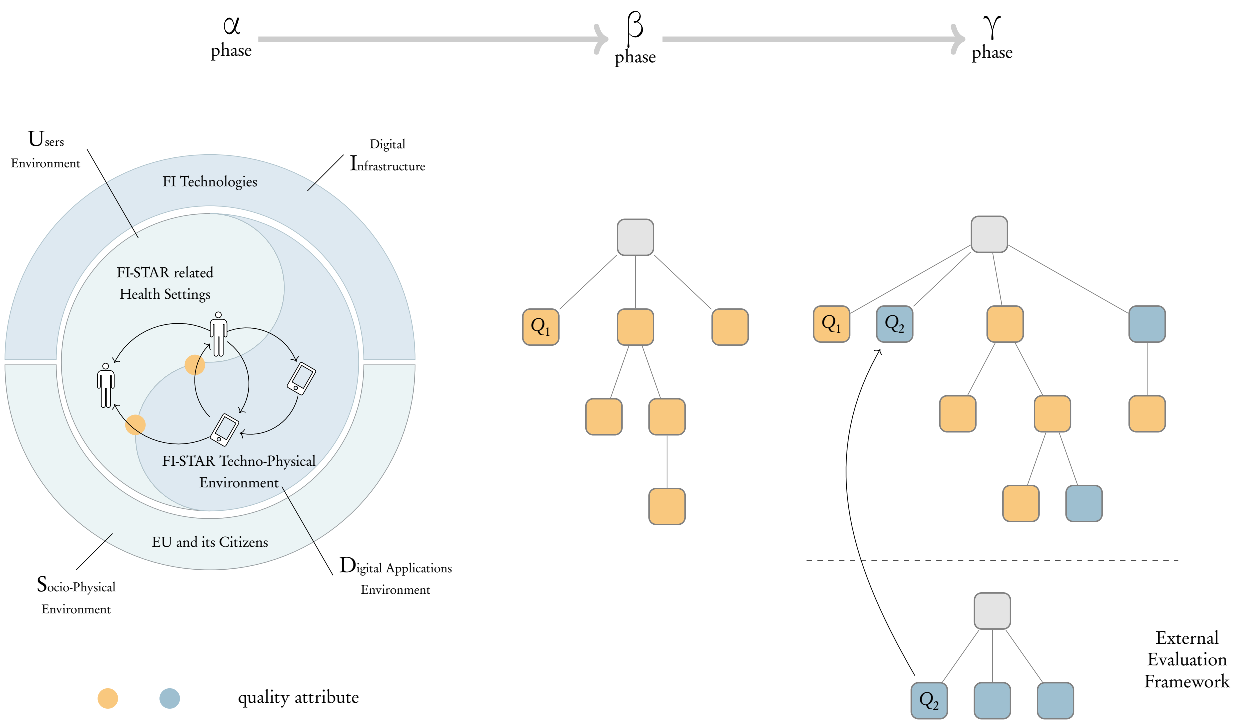

Figure 2: Ontology construction for a health information system

The γ phase ontology is constructed the same as the β phase but it adds some more materials (quality attributes) from external sources. In this sense, the quality attributes specified in an external evaluation framework, probably a model-based one, should be extracted first. Those quality attributes should be fed into the β ontology the same as other quality attributes during the β phase. The UVON method does not discriminate between quality attribute by the origin, but it might be a good practice to mark those quality attributes originally from the external evaluation framework if we need later to make sure they are used by their original names in the summarizing level (to be discussed in the following paragraphs).

EACH level of the resulted ontology tree(s) —except those that are deeper than the length of the shortest branch— represents or summarizes quality attributes of the whole system in some degree of generality or specificness. That is the *root* node is the most general quality attribute, which is too general to be useful for any evaluation, and the levels below, each gives a view of the quality attributes in the whole systems. As each parent node represent a general form of its children, therefore each level summarizes the level below. We refer to one of these levels of the ontology tree that is considered for creating the questionnaire as the *summarizing level*.

the summarizing level

The quality attributes in each of other levels (such as L_1 in Figure 4) can be evaluation aspects or the answer to *what to evaluate* that can be measured by a questionnaire or other measurement methods. Also, depending on the measuring method, the level below the summarizing level can be used for detailing each of the evaluation aspects.

The practicalities of measurement in a case determines which summarizing level to choose. Levels closer to the root can be too abstract while deeper levels can be too much detailed. Also, the number of quality attributes in a level can impact which level is appropriate. In the FI-STAR project, the limitation on the number of questions in the questionnaire was a determinant for selecting the summarizing level, where only level 2 was fitting in the project limitations; although level 3 helped to make each question more detailed. It is possible to grow a short branch by adding a chain of children that are the same as their parents and make the branch reach a specific level, hence being able to select that level as a summarizing level.

3.2 Result of the UVON Method Application in the FI-STAR Project

HARVESTING the value-cases and requirement documents for all seven trial-cases in the FI-STAR project provided the initial set of quality attributes, i.e. the α set. Several quality attributes were redundant or similar, but it was left to the UVON method to unify them. There were also several quality attributes with the same wording, but different conceptual indications in their specific context of usage. These quality attributes we added to the α with small modifications to differentiate them from each other . For example, two different references to *efficiency* were converted to *efficiency by reducing complexity* and *efficiency by reducing time*.

In the next step, i.e. β phase, the UVON method developed β ontology by using the α set. The redundant quality attributes were integrated into single entities, while other quality attributes grouped by their direct or indirect parents in the ontology structure regarding their degree of similarity or dissimilarity.

Also, it was noticed that quality attributes are preferred —though, not always necessarily— to be noun phrases rather than adjective phrases; because fulfilling a quality attribute expressed in an adjective phrase could imply that all of its child quality attributes need to be fulfilled. For example, to fulfill the quality of being *safe*, it is required to be both *safe for patient* and *safe for medical personnel*. This is in contrast to the *child is type of parent* relations that we considered between the ontology entities. But, if we consider the noun form (noun phrase), for example to consider *safety* rather than *safe*, then *safety for patient* and *safety for medical personnel* are all subtopics of *safety* hence that would be correct and more intuitive. Also, considering that each node in the ontology is an aspect for evaluation can make deciding parent-child relations more straightforward. So, for example, the *safety* node should be read as *safety aspect* and its child should be read as *safety for patient aspect*

Applying the UVON method in its β and γ phases, respectively created the β and γ ontology structures. The first ontology structure (β) is based on the α set of collected quality attributes; where the second one (γ) extends the β ontology by integrating the MAST framework evaluation aspects (grouped as domains) as specified by MAST (Kidholm et al., 2012). Here, 'integration is the process of building an ontology in one subject reusing one or more ontologies in different subjects' (Pinto and Martins, 2001). In this sense, γ ontology is constructed by mapping, aligning, or merging (Abels et al., 2005) the ontological representation of the external framework evaluation aspects (MAST in our case) to the β ontology. The result of the integration is shown in Table 2.

The MAST framework specifies 7 evaluation domains, where each contains several topics (aspects or sub-aspects) (Kidholm et al., 2012). Due to the FI-STAR project requirements, we ignored *clinical effectiveness* and the *Socio-cultural, ethical, and legal* domains (job of other teams). One another domain, *health problem and description of the application*, and some of the aspects in other domains could not be considered as quality attributes and were removed from the process. The remaining four domains that were fed into the UVON method are safety, patient perspectives, economic aspects, and organizational aspects. There was an interesting observation, probably a motivation for further investigations, that the aspects in those four domains overlap considerably with the evaluation aspects that were elicited from the FI-STAR users and were formed as an ontology by the UVON method.

The final ontology structure, i.e. γ, has 400 nodes (a two significant figures number) originated from the β ontology (which itself is originated from the quality attributes gathered in the α set) or the mast framework. Reports on the number of nodes in β or γ ontologies should be considered with care, as it might have a limited number of significant digits. This is due to partial subjectivity in the method, as described before, so the possibility of reaching different numbers in each run of the method. Regarding the observation during developing the ontologies for the seven trial cases of

FI-STAR, it is expected that new future cases, of similar nature, would not add much increase in the number of nodes, but many of them use the same or similar quality attributes (nodes).

Both the β and γ ontology structures were described in Web Ontology Language (OWL) using Prot version 4.x software. Web Ontology Language (OWL), as an ontology language, can describe a domain of knowledge through its lingual elements and their relations (Bechhofer, 2009). In OWL, there exist *individuals*, *classes*, *class relations*, *individual relations*, and *relation hierarchies* (Hitzler et al., 2012). In FI-STAR ontology structures, the individuals were mapped to the use-cases in the FI-STAR project; classes were used to represent quality attributes (i.e. the evaluation aspects); and the class relations became the hierarchal relation between quality attributes (i.e. *is of type* or the super-class to sub-class relation). Individual relations and relation hierarchies were not utilized.

Some generic nodes were inserted in order to group sibling nodes that were conceptually closer together in the ontology structure. If a quality attribute was connected to two different branches, it was forked and presented in the both branches (as described before); that keeps the ontology in a tree structure rather than an acyclic directed graph.

Quality Attributes	
1. Accessibility	6. Efficiency
2. Adhereability	7. Effectiveness
3. Affordability	8. Empowerment
4. Authenticity	9. Safety
5. Availability	10. Trustability

Table 1: The list of quality attributes appeared in the second level of the ontology by using the UVON method in the FI-STAR project.

APPLYING the UVON method in the FI-STAR project case, at the end of the method γ phase, 10 nodes appeared below the root of the ontology tree (Table 1). These 10 quality attributes at the second level of the tree are parents to other child nodes, hence each is the unification and aggregation of other quality attributes that are originated either from the FI-STAR requirement documents or the MAST framework and reside below these 10 quality attributes. The number 10 was in the scope of practical considerations for creating an evaluation questionnaire for the FI-STAR project, but we also considered the third level of the tree to provide more details for each question in the questionnaire. Due to separation of responsibilities in the FI-STAR project, these 10 quality attributes do not represent some aspects such as the *clinical effectiveness* or *legal and ethical* ones, hence, the number 10 would be larger if we had included those aspects when applying the UVON method in the project.

In the FI-STAR project, the measurement of evaluation aspects is performed through a questionnaire that is based on those 10 extracted aspects

Table 2: The mapping between MAST evaluation aspects and the final evaluation aspects for the FI-STAR project through using UVON

MAST		Final top aspect
Domains	Aspects	
Health problem and description of the application		*
Safety		
	Clinical safety (patients and staff)	Safety
	Technical safety (technical reliability)	Safety
Clinical effectiveness		**
	Effects on mortality	**
	Effects on morbidity	**
	Effects on health related quality of life (HRQL)	**
	Behavioural outcomes	** (but can relate to Adhereability)
	Usage of health services	** (but can relate to Adhereability)
Patient perspectives		
	Satisfaction and acceptance	***
	Understanding of information	Accessibility
	Confidence in the treatment	Trustability and Authenticity
	Ability to use the application	Accessibility
	Access and accessibility	Accessibility
	Empowerment, self-efficacy	Empowerment
Economic aspects		
	Amount of resources used when delivering the application and comparators	Efficiency
	Prices for each resource	Efficiency
	Related changes in use of health care	*
	Clinical effectiveness	**
	Expenditures per year	Affordability
	Revenue per year	**
Organizational aspects		
	Process	* (but can relate to Efficiency)
	Structure	*
	Culture	*
Socio-cultural, ethical, and legal aspects		**

* Not a quality attribute

** Not included because of the FI-STAR project definition and division of tasks

*** Had been already covered by some generic questions in the questionnaire

L_1	Does the application increase *efficiency* by reducing ...	Strongly Agree	Agree	No Idea	Disagree	Strongly Disagree
L_2	Complexity or number of tasks?					
	Number of reworks?					
	Time consumed?					

Figure 3: Sample questionnaire output from the UVON method

in the γ ontology. In practice, the questionnaire found two versions of the same topic but different wordings, one for the patients and the for health professionals. In practice, the questionnaire was formed in two versions, one for the patients and one for the health professionals, where each expresses the same concept in two different wordings (an operation theatre case did not have patient questionnaire).

Generally and regarding practicalities of an evaluation case, it is possible to consider deeper levels of the resulted γ ontology in that specific case. In the FI-STAR case, this possibility is reflected in a sample question on *efficiency* from the questionnaire (Figure 3), where a general question got more detailed by considering other quality attributes below in the third level of the ontology. This possibility of going deeper is also depicted in Figure 4.

In the FI-STAR project, the quality attributes (and later the questionnaires) were delivered to each case's stakeholders and it was asked to validate the relevancy of each quality attribute or the corresponding question regarding their case. All the cases in the FI-STAR validated and approved the relevancy of those, while some asked for minor changes in the wordings of some of the questions to be more clear for the patient respondents in their case.

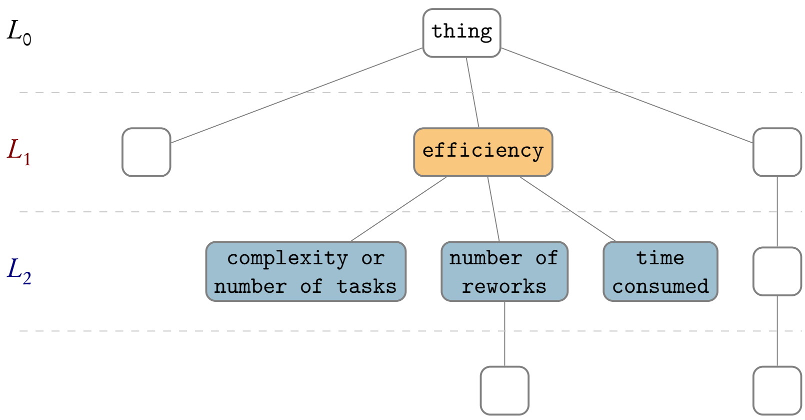

Figure 4: More details can be evaluated by looking at deeper nodes in the ontology structure

4 *Discussion*

what are ontologies

ONTOLOGIES are the formal and computable ways of capturing knowledge in a domain —whether local or global (Uschold, 2000)— by specifying the domain key concepts (or objects) and interconnecting them by a predefined set of relations (Noy and McGuinness, 2001). The formality and computability help to communicate knowledge between people or software agents, enable reuse of knowledge, make explicit declaration of the assumptions, and facilitate the analysis and study of the domain knowledge (Noy and McGuinness, 2001). Inference algorithms can infer and extract new knowledge or predict or deduct new situations by working on an ontology. As it is reflected in the above ontology description, an ontology is structured as a network (mathematically a graph). Limiting the kind of relations between the concepts might result to specific structural forms such as trees.

An ontology would be formed as hierarchy if the relations between the concepts are limited to the *is of type* relation, where each none-leaf concept is a more generic form or super-class to its children. This hierarchy can be an *acyclic direct graph* if we allow one concept to be a sub-class of more than one other concepts, and it would be a *tree* if one concept is sub-class of only one other concept. The acyclic directed graph can be converted to a tree if we replicate the same concept-leaf in different branches. The unification that exists in the nature of a tree graph, i.e. unification of branches toward the root, is the source of unification that we want to apply for the evaluation of quality attributes in the health information systems and that is why the UVON method create this type of structure.

unification nature of tree ontologies

Ontologies are traditionally the output of manually content curation and its associated consensus establishment processes (Alterovitz et al., 2010). Nevertheless automated or semi-automated methods of ontology construction might expose considerable advantages in efficiency, repeatability, and uniformity. The UVON method described in this paper, uses a semi-automatic approach toward creating tree-style ontologies for the sake of extracting evaluation aspects.

automatic, semi-automatic, and manual ontology creation

4.1 Extending the Evaluation Utilizing the Ontology

The ontological representation of a health information ecosystem, gives a computable structure, from which several indications, including evaluation aspects, can be extracted. Functions can be defined on this ontology that quantify, combine, compare, or select some of the nodes or branches. The ontology itself can be extended by assigning values to its nodes and edges. The ontology gives the possibility of further inferences. For example, if two nodes (quality attributes) are disjoint (mutually exclusive) any two child from each of them would be disjoint respectively. If during the application of the UVON method, by mistake, one quality attribute gets replicated into two disjoint branches then this mistake can be detected and avoided automatically (replication would be disallowed between those specific nodes).

digging the enriched ontologies for more indications

As discussed in Section 3.2 and shown in Table 2, we skipped the *clinical effectiveness* and *socio-cultural, ethical, and legal* domains from the MAST framework due to the project definition. Nevertheless,the UVON method can consider those aspects when it is applicable and there is no project restrictions. Therefore, we hope to witness more inclusive application of the UVON method in the future cases.

Also, the selection of the MAST framework was due to the common theme of the e-health applications in the FI-STAR project. We encourage application of the UVON method by considering relevant evaluation frameworks, not necessarily MAST. The results of those applications can demonstrate the powers, weaknesses, and extension points of the FI-STAR method.

The UVON method is context-insensitive by its approach. Still, more empirical evidences with a higher degree of diversity are need to examine what are the challenges or advantages of applying the UVON method in a more diverse range of fields beyond health information systems.

context insensitive

4.2 *Limitations of the UVON Method*

The UVON method is subjected to conceptual and methodological limitations in its capacities. Probably a prominent conceptual limitation is the fact that the method does not represent or give an account of the dynamics of the health information systems; hence cannot facilitate their evaluation. The relations in the UVON constructed ontologies are restricted to the *is of type* relationship and cannot reflect how qualities or other indicators impact each other. The absence of insight about the dynamics of a health care information system prevents predictive evaluations. In consequence, any emergent behavior that is not explicitly captured by requirement documents or the to-be-merged external evaluation framework, is going to be ignored. From the other side, still it can be imagined that the output ontologies of the UVON method can be used as scaffolds in models that incorporate dynamics of health care information systems.

no consideration of dynamics

The UVON method partially relies on subjective decision makings that can create methodological limitations and challenges. Although the main strategy in the UVON method is to minimize these subjective decision makings, still the existing ones can result in creating different ontologies by different applicants of the method. As a suggestion, for the sake of reaching more convergence, it is possible to think of enhancing the method with more objective lexical analytics methods. Methods of ontology construction and integration, especially those concerning class inheritance analysis (Abels et al., 2005), can be valid candidates for these types of methods.

partial subjectivity

UVON generated ontologies are not advised for universal application. However, for a new case of evaluation, a UVON-generated ontology that was developed for similar cases can be considered as an alternative to developing a new ontology regarding the project resource limitations. This reuse should be accomplished with this consideration that quality attributes of the same wording might indicate slightly different meanings in a different case. This case-sensivity of meanings might result in different sub-class and super-class relations, changing the structure of the ontology and making the reuse of ontology unadjusted.

context specificness

The UVON method cannot guarantee that in the output ontology all the branches that begin from the root would reach to the level of the tree (i.e. have a node in that level) that we want to base our questionnaire (or any other measurement method). Hence, a short branch needs to get extended to appear in some specific tree level and the questionnaire that is based on that tree level. Also, the method does not guarantee that the quality attributes in that level are all of the same degree of generality of specificness. It is also not gauranteed the number of nodes (quality attributes) in any level matches the practicalities of evaluation; they can be too few or too many. For example, in the FI-STAR case, the number of quality attributes in the target level (level 2) had to match with the appropriate maximum number of questions that could be put in a questionnaire and fortunately it was within that boundaries.

It is also possible, at least in theory, that all quality attributes, end up being a direct child of the root *thing* node; where this dwarf and horizontally inflated ontology structure does not unify any of the child quality attributes; hence the method output would be useless.

The methodological limitations can cause the need for manual adjustments, such as adding extra nodes in between some parent-child nodes. Of course, the manual adjustments can add more subjectivity into the formation of the ontologies.

The UVON method permits integrating evaluation aspects from other evaluation frameworks. Still, it does not guarantee that the result would include all features of the integrated evaluation framework. Still, this integration is about the suggested evaluation aspects by those evaluation frameworks. If a framework dynamically changes its suggested evaluation aspects, for example based on the evaluation case specifications, the UVON does not follow that dynamic feature. Also, the straightforward wordings for an evaluation aspect in an evaluation framework might get obscured by going through the integration process in the UVON method, being replaced by more generic terms.

is integration always better?

5 *Conclusion*

The unifying nature of ontologies, when they are in tree form, can be used to create a common ground of evaluation for heterogeneous health technologies. Ontologies can be originated from requirement and value-case documents, i.e. being internal, or they can be extracted from available external evaluation frameworks, i.e. being external, or being originated from a mix of both internal and external sources. The UVON method, introduced in this paper, was able to create a common ground for evaluation by creating an ontology from requirement and value-case documents of the seven trial projects in the FI-STAR project, and extending that ontology by mixing elements from the MAST evaluation framework. The UVON method can be used in other similar cases to create ontologies for evaluation and mix it with elements from other evaluation frameworks.

The UVON method is in contrast to other methods that do not consider case-specific internal requirements or cannot easily get extended to include other evaluation frameworks. The ontological structure of evaluation aspects created by the UVON method offers the possibility of further investigations for other indications related to evaluation of the subject systems.

The final result of applying the UVON method in the FI-STAR project resulted in 10 evaluation aspects to be chosen for measurement. This set of evaluation aspects can grow adaptive to project changes, be repeated in similar cases, and be a starting point for future evaluations in similar projects. By applying the UVON method in more cases, a possible stable result can be suggested as the set of generic evaluation aspects that are usable in evaluation cases similar to FI-STAR.

6 *Acknowledgment*

The authors would like to acknowledge the contribution of project partners from the FI-STAR project for providing the context of this study. The FI-STAR project is funded by the European Commission under the Seventh Framework Programme (FP7), under grant agreement number 604691.

Regarding the contributions, SE drafted the paper, incorporated contributions from other authors into the paper, contributed to the design of study, developed the proposed model and method, processed data for β and γ phases of the proposed method, and contributed to the proposed method final result. PA contributed to design of the study, contributed to the proposed method final result, supervised the research process, and reviewed and commented on the paper. TL contributed to design of the study, supervised the research process, and reviewed and commented on the paper. SF collected data for the α phase of the proposed method, and reviewed and commented on the paper. JB contributed to design of the study, supervised the research process, and reviewed the paper.

7 *Conflict of Interests*

None declared.

8 *References*

References

Abels, S., L. Haak, and A. Hahn (2005). Identification of Common Methods Used for Ontology Integration Tasks. In *Proceedings of the First International Workshop on Interoperability of Heterogeneous Information Systems*, IHIS '05, New York, NY, USA, pp. 75–78. ACM.

Alterovitz, G., M. Xiang, D. P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M. A. Harris, M. E. Dolan, J. A. Blake, and M. F. Ramoni (2010, February). Ontology engineering. *Nature Biotechnology 28*(2), 128–130.

Ammenwerth, E., J. Brender, P. Nykänen, H.-U. Prokosch, M. Rigby, and J. Talmon (2004, June). Visions and strategies to improve evaluation of health information systems: Reflections and lessons based on the HIS-EVAL workshop in Innsbruck. *International Journal of Medical Informatics 73*(6), 479–491.

Ammenwerth, E., S. Gräber, G. Herrmann, T. Bürkle, and J. König (2003, September). Evaluation of health information systems—problems and challenges. *International Journal of Medical Informatics 71*(2–3), 125–135.

Ammenwerth, E., C. Iller, and C. Mahler (2006, January). IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. *BMC Medical Informatics and Decision Making 6*(1), 3.

Bechhofer, S. (2009). OWL (Web Ontology Language). In L. LIU and M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems*, pp. 2008–2009. Springer US. bibtex: bechhofer_owl_2009b.

Berg, M. (2001, December). Implementing information systems in health care organizations: myths and challenges. *International Journal of Medical Informatics 64*(2–3), 143–156.

Busse, R., J. Orvain, M. Velasco, M. Perleth, M. Drummond, F. Gürtner, T. Jørgensen, A. Jovell, J. Malone, A. Rüther, and C. Wild (2002). BEST PRACTICE IN UNDERTAKING AND REPORTING HEALTH TECHNOLOGY ASSESSMENTS. *International Journal of Technology Assessment in Health Care 18*(02), 361–422.

Bürkle, T., E. Ammenwerth, H.-U. Prokosch, and J. Dudeck (2001). Evaluation of clinical information systems. What can be evaluated and what cannot? *Journal of Evaluation in Clinical Practice 7*(4), 373–385.

Chaudhry, B., J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle (2006, May). Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care. *Annals of Internal Medicine 144*(10), 742–752.

Cheng, B. H. C. and J. M. Atlee (2007). Research Directions in Requirements Engineering. In *2007 Future of Software Engineering*, FOSE '07, Washington, DC, USA, pp. 285–303. IEEE Computer Society.

Choi, N., I.-Y. Song, and H. Han (2006, September). A Survey on Ontology Mapping. *SIGMOD Rec. 35*(3), 34–41.

Cusack, C. M. and E. G. Poon (2007). Health information technology evaluation toolkit. *Rockville, MD: Agency for Healthcare Research and Quality*.

Davis, F. D. (1989, September). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly 13*(3), 319–340.

DeLone, W. H. and E. R. McLean (2004, October). Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model. *International Journal of Electronic Commerce 9*(1), 31–47.

Draborg, E., D. Gyrd-Hansen, P. Bo Poulsen, and M. Horder (2005). International comparison of the definition and the practical application of health technology assessment. *International Journal of Technology Assessment in Health Care 21*(01), 89–95.

Ekeland, A., A. Bowes, and S. Flottorp (2012). Methodologies for assessing telemedicine: A systematic review of reviews. *International Journal of Medical Informatics 81*(1), 1–11.

Facey, Karen, Topfer, Edited Leigh-ann, Chan, Liza, and Facey, Karen (2006). INAHTA Health Technology Assessment (HTA) Glossary. http://www.inahta.org/Glossary/.

FISTAR Consorium (2014). About FI-STAR. https://www.fi-star.eu/about-fi-star.html.

Friedman, C. P. and J. Wyatt (2006). *Evaluation methods in biomedical informatics*. Springer Science & Business Media.

Goodhue, D. and R. Thompson (1995). Task-technology fit and individual performance. *MIS Quarterly: Management Information Systems 19*(2), 213–233.

Greenhalgh, T. and J. Russell (2010, November). Why Do Evaluations of eHealth Programs Fail? An Alternative Set of Guiding Principles. *PLoS Med 7*(11), e1000360.

Hitzler, P., M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph (2012). OWL 2 web ontology language primer. *W3C recommendation 27*(1), 123.

Holden, R. J. and B.-T. Karsh (2010, February). The Technology Acceptance Model: Its past and its future in health care. *Journal of Biomedical Informatics 43*(1), 159–172.

Hu, P., P. Chau, O. Liu Sheng, and K. Tam (1999). Examining the Technology Acceptance Model Using Physician Acceptance of Telemedicine Technology. *Journal of Management Information Systems 16*(2), 91–112.

Kidholm, K., A. G. Ekeland, L. K. Jensen, J. Rasmussen, C. D. Pedersen, A. Bowes, S. A. Flottorp, and M. Bech (2012). A MODEL FOR ASSESSMENT OF TELEMEDICINE APPLICATIONS: MAST. *International Journal of Technology Assessment in Health Care 28*(01), 44–51.

Kreps, D. and H. Richardson (2007, July). IS Success and Failure—The Problem of Scale. *The Political Quarterly 78*(3), 439–446.

Kristensen, F. B. (2009, June). Health technology assessment in Europe. *Scandinavian Journal of Public Health 37*(4), 335–339.

Lampe, K., M. Mäkelä, M. V. Garrido, H. Anttila, I. Autti-Rämö, N. J. Hicks, B. Hofmann, J. Koivisto, R. Kunz, P. Kärki, A. Malmivaara, K. Meiesaar, P. Reiman-Möttönen, I. Norderhaug, I. Pasternack, A. Ruano-Ravina, P. Räsänen, U. Saalasti-Koskinen, S. I. Saarni, L. Walin, and F. B. Kristensen (2009). The HTA Core Model: A novel method for producing and reporting health technology assessments. *International Journal of Technology Assessment in Health Care 25*(Supplement S2), 9–20.

Littlejohns, P., J. C. Wyatt, and L. Garvican (2003, April). Evaluating computerised health information systems: hard lessons still to be learnt. *BMJ 326*(7394), 860–863.

Noy, N. F. and D. L. McGuinness (2001). *Ontology development 101: A guide to creating your first ontology*. Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880.

Pinto, H. S. and J. P. Martins (2001). A methodology for ontology integration. In *Proceedings of the 1st international conference on Knowledge capture*, pp. 131–138. ACM.

Talmon, J., E. Ammenwerth, J. Brender, N. de Keizer, P. Nykänen, and M. Rigby (2009, January). STARE-HI—Statement on reporting of evaluation studies in Health Informatics. *International Journal of Medical Informatics 78*(1), 1–9. bibtex: talmon_stare_2009.

Tarjan, R. (1972, June). Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing 1*(2), 146–160.

Uschold, M. (2000). Creating, integrating and maintaining local and global ontologies. In *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000)*. Citeseer.

Yusof, M. M., J. Kuljis, A. Papazafeiropoulou, and L. K. Stergioulas (2008, June). An evaluation framework for Health Information Systems: human, organization and technology-fit factors (HOT-fit). *International Journal of Medical Informatics 77*(6), 386–398.

Yusof, M. M., A. Papazafeiropoulou, R. J. Paul, and L. K. Stergioulas (2008, June). Investigating evaluation frameworks for health information systems. *International Journal of Medical Informatics 77*(6), 377–385.