Ontological approach for semantic modeling and querying the Qur'an

Aimad Hakkoum[1], **Said Raghay**[2]
[1]Cady Ayyad University, Marrakesh, Morocco
[2]Cady Ayyad University, Marrakesh, Morocco
[1]imad.hakkoum@ced.uca.ma, [2]s.raghay@uca.ma

ABSTRACT

The Qur'an is considered as the first source of knowledge and guidance for Muslims throughout the world. It's a difficult book to understand and to interpret without consulting domain experts and Qur'anic commentaries (Tafsir books). Therefore it's very sensitive to analyze and to model his content for fear to make bad assumptions and axioms. In recent years a number of researches has been done to facilitate the retrieval of knowledge from the Qur'an, but most of the available researches are using human readable data resources and therefore cannot be reused and linked using semantic web technologies. This is why in this project we will adopt an approach that enables humans and computers to understand the Qur'an knowledge throughout the creation of a Qur'anic ontology. The goal of the ontology is to build a computational model capable of representing as much as possible of the concepts mentioned on the Qur'an and the relationships between them using Protégé-OWL. The ontology can be queried using SPARQL queries. For non-technical users we will build a tool that enables them to browse the content of the ontology.

***Keywords*: Qur'an, semantic web, ontology, knowledge extraction.**

1. INTRODUCTION

Ontology is one of the emerging specialty of research in computer science and semantic web. It can be defined as «an explicit specification of a conceptualization» (Gruber, 1995). Ontologies explicitly structure and represent domain knowledge in a machine-readable format so they can be incorporated into computer-based applications and systems to facilitate automatic annotation of web resources, domain representation and reasoning task, decision support, and natural-language processing so as to serve as an integral part of the Semantic Web (Shadbolt et al., 2006).

The Qur'an is the religious text of Islam, revealed to the Prophet Muhammad and distinguished by its miraculous style. It is considered as the basic reference for all Islamic sciences and, in fact, of the Arabic language. Therefore the Qur'an remains in the eyes of Muslims a unique book of its kind that deserves learning, studying and preserving. In recent years the Qur'an was the subject of numerous researchers in the field of computer science. Most of them use taxonomy, hierarchy or tree structure to present and classify the Qur'an knowledge. These approaches are still effective to answer most of user's queries but cannot be reused and linked using web semantic technologies. This is why in this project we will adopt an approach that enables humans and computers to understand the Qur'an knowledge throughout the creation of a Qur'anic ontology. The ontology will be created using Protégé[1]. We will also use Jena Framework[2] and Jena triple database (TDB) to manipulate and query the ontology. Both of these tools support Arabic language to write and display RDF data0 (Beseiso et al., 2010).

We will start by presenting the existing work done so far in the knowledge representations of the Qur'an and theology in general. We will focus on the researches that are going to be used to achieve our task. After that we will discuss the methodology used to extract and model the content of the concepts mentioned on the Qur'an and the relationships between them. Finally we will address what is left to be done in this project.

2. RELATED WORK

Ontology-Based Researches

Semantic web technology is still lacking a critical mass of RDF data online and up-to-date terms. Ontologies are missing for many application domains especially for Islamic sciences. In order to address this lack, the Qur'an became in recent years a target of interest for studies in the field of Semantic web technologies. Because of the complexity of the task and the necessary time to extract all the knowledge contained in the Qur'an, several researchers tried to cover a specific topic from the Qur'an like prayer (Salat)0(Saad et al., 2010), faith (Iman) and deed (Akhlaq) 0(Ta'a et al., 2013) or Umrah 0(Sharef et al., 2013).

One of the important ontologies developed so far in the subject is the Semantic Qur'an dataset (Sherif et al., 2014). The namespace of the ontology is QVOC (Qur'an vocabulary) which consists of a multilingual RDF representation of translations of the Qur'an. The resulting RDF data encompasses 43 different languages which belong to the most under represented languages in Linked Data, including Arabic, Amharic and Amazigh. The Semantic Qur'an dataset is published at http://datahub.io/dataset/semanticquran. It contains over 15 million RDF triples, the Qur'an words are linked to 7718 word from dbpedia and 18655 from Wiktionary.

Another useful project is the Qur'anic Arabic Corpus[3] (QAC) and discussed in Kais Dukes PhD Thesis (Dukes, 2013). It's an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Qur'an. It contains also an ontology of 300 concepts with 350 relations mainly of type "Instance Of". It was developed using Knowledge Interchange Format (KIF). This ontology was translated to OWL and enhanced by designing more relationships and restrictions using sources from the Qur'an, hadith, Islamic websites and other Islamic related resources 0(Aliyu et al., 2013).

Text mining researches

The web contains a lot of resources that provide the Qur'an text and other Islamic books in different formats (HTML, Text, PDF, SQL dump and XML). It enables to do keyword search in an advanced way using lemmas, roots, word proximity and Boolean search. These resources are available in different languages especially in Arabic and English.

[1] http://protege.stanford.edu/

[2] http://jena.apache.org/

[3] http://corpus.quran.com

2.1.1 Tanzil Project

The Tanzil Project[4] was launched in early 2007 with the aim of producing a curated Unicode version of the Arabic Qur'an text that can serve as a reliable standard text source on the web. To achieve this goal, Tanzil team subsequently developed a three-step data quality assurance pipeline; which consisted of (1) an automatic text extraction of Arabic Qur'an text, (2) a rule-based verification of the Arabic Qur'an text against a set of grammatical and recitation rules and (3) a final manual verification by group of experts.

The Qur'an text from Tanzil project is widely used by a number of web sites and research groups, and it was validated by different entities like "King Fahad Qur'an Complex"[5].

2.1.2 Qur'anic Arabic Corpus

The Qur'anic Arabic corpus (QAC) provide also an important dataset that contains the morphology structure of each word in the Holy Qur'an. This corpus was produced using Buckwalter morphology analyzer followed by manual checking. It gives others useful morphological features like gender information verb forms and plurality.

2.1.3 The Qur'an Annotation for Text Mining

This resource[6] contains several useful tools to understand the Qur'an; especially two important ones that we will use in our work. The first one is QurAna0(Sharaf and Atwell, 2012a) which annotate the antecedent of every pronoun in the Qur'an. It relies on QAC to extract all the pronouns from the Qur'an then a Manuel annotation had been performed, and finally the result was put on the QAC website for further validation by users. The results are available at www.textmining.com

The other valuable tool from this project is QurSim (Sharaf and Atwell, 2012b) which provides a dataset of related verses. It was based not only on common words or roots but also on Ibn Kathir commentary (Tafsir) of the Qur'an where he cited some relative verses when commenting on a verse. After extracting related verses from Ibn Kathir commentary a manual check was done to class the result into 3 degrees of relatedness: loosely related to strongly related.

2.1.4 Qurany

This research (Abbas, 2009) proposes a tool7 that categorizes the topics discussed in the Qur'an verses to a comprehensive index that covers nearly 1100 topics in the Qur'an. It classifies the Qur'an into fifteen main themes and subdivides the main themes into sub themes and sub sub themes and so on.

[4] http://tanzil.net/wiki/Tanzil_Project

[5] http://www.qurancomplex.org/

[6] http://www.textminingthequran.com/wiki/Main_Page

[7] http://quranytopics.appspot.com

3. ONTOLOGY DEVELOPMENT

At the present time there is no consensus on the best practices to follow when developing an ontology. There are more than 33 methods of ontological engineering0(Psyché et al., 2003). There is obviously no method which is the best. However, whatever method is adopted, it must refer to the fundamental rules in ontology design, which are:

- There is no one correct way to model a domain. The best solution always depends on the application that you have in mind and the extensions that you anticipate.
- Ontology development is necessarily an iterative process.
- Concepts in the ontology should be close to objects and relationships in the domain of interest. These concepts should reflect the model as in the real world.

We are going to follow the methodology discussed in 0(Noy, 2001) by adopting an iterative approach to ontology development with the six steps: define the ontology domain and scope, review existing ontologies, enumerate important terms in the ontology, define the classes and the class hierarchy, define the properties of classes and their facets and finally create instances.

Domain and scope

The ontology will cover the Qur’an knowledge, the ontology must allow semantic indexing of the Qur'anic content and the relation between the extracted concepts.

We will cover the following subjects: Qur’anic chapters and verses, each word of the Qur’an and its root and lemma to facilitate key word search. We will not cover words morphology but we will add links to QVOC ontology where this is covered; however we will cover the pronouns in order to define their antecedents.

Ontology reuse

We are going to reuse the two Qur’anic ontologies: Semantic Qur’an (QVOC) and QAC ontology with the OWL format. We can see ontology reuse according to two different points of view: building an ontology, by assembling, extending, specializing and adapting, other ontologies, or building an ontology, by merging different ontologies on the same or similar subject into a single one that unifies all of them (Pinto and Martins, 2000).

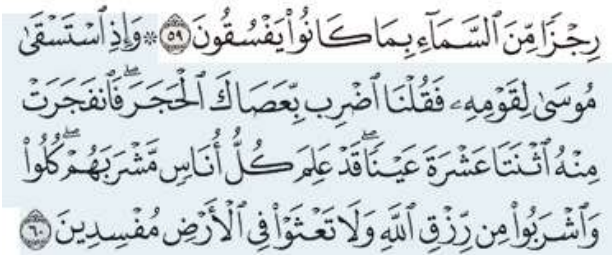
The first kind of reuse is named *ontology integration*. The second kind of reuse is named *ontology merge*, we will use the first method because to use the second method all of the reused ontologies must be always available and highly maintained which is not true in our case; however to keep the link with the QVOC resources we are going to use the OWL property “owl:SameAs” to state that the two resources represent the same thing and they can be interchangeable.

Enumerate important terms in the ontology

As we already mentioned the Qur’an text is very complex therefore we cannot rely on automatic extraction to detect the important terms that we must include in the ontology, but instead we must rely on the understanding of each verse using one or more commentary books to extract all the information contained in the verse explicitly or implicitly.

There are two approaches to create the Qur'an ontology: verse by verse extraction and topic extraction.

In verse by verse extraction we have to analyze each verse and build the ontology progressively in a linear way. This will be an incredibly time consuming process. The following example demonstrate how we can extract information from the verse 2:60 (chapter 2, verse 60):



رِجْزًا مِّنَ ٱلسَّمَآءِ بِمَا كَانُوا۟ يَفْسُقُونَ ﴿٥٩﴾ ۞ وَإِذِ ٱسْتَسْقَىٰ
مُوسَىٰ لِقَوْمِهِۦ فَقُلْنَا ٱضْرِب بِّعَصَاكَ ٱلْحَجَرَ ۖ فَٱنفَجَرَتْ
مِنْهُ ٱثْنَتَا عَشْرَةَ عَيْنًا ۖ قَدْ عَلِمَ كُلُّ أُنَاسٍ مَّشْرَبَهُمْ ۖ كُلُوا۟
وَٱشْرَبُوا۟ مِن رِّزْقِ ٱللَّهِ وَلَا تَعْثَوْا۟ فِى ٱلْأَرْضِ مُفْسِدِينَ ﴿٦٠﴾

Fig 1: verse 2:60 from Ayat project (http://quran.ksu.edu.sa)

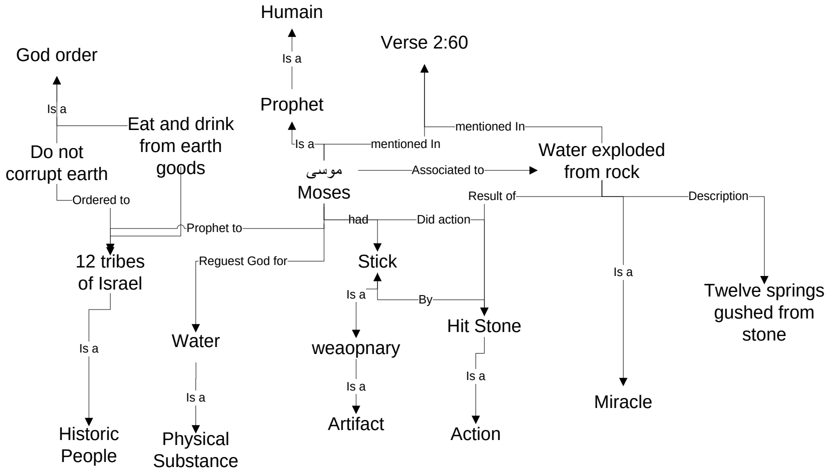
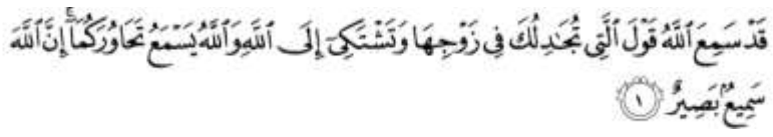


Fig 2: verse 2:60 knowledge extraction

This approach requires that we cover all the Qur'an otherwise the resulting model will be incoherent because we will cover a fragment of each subject and it won't be very useful. The other approach is to cover only some topics by only analyzing their related verses. This way after adding a topic, the ontology will be in a coherent form and can be used and published.

We will use the second approach with the following topics: chapters, verses, words, verse topics, pronouns antecedents, people, events and places cited on the Qur'an. To extract these concepts we will use a manual approach because of the complexity of the Qur'anic text using the previously cited research and especially the Tafsir books. Here is an example of extracting the list of persons cited on the Quran for the verse 58:1 (chapter 58, verse 1):



قَدْ سَمِعَ ٱللَّهُ قَوْلَ ٱلَّتِى تُجَٰدِلُكَ فِى زَوْجِهَا وَتَشْتَكِىٓ إِلَى ٱللَّهِ وَٱللَّهُ يَسْمَعُ تَحَاوُرَكُمَآ ۚ إِنَّ ٱللَّهَ
سَمِيعٌۢ بَصِيرٌ ﴿١﴾

Fig 3: verse 58:1

This verse refers to Khaulah bint Tha'labah (خولة بنت ثعلبة) and her husband Aws bin As-Samit (أوس بن الصامت), we can see that these two persons are mentioned implicitly in this verse and are not mentioned anywhere else in the Qur'an.

Define the classes and the class hierarchy

We used Protégé and OWL to create the ontology because it's well maintained and contains a number of useful plugin that we can add to facilitate reasoning tasks and visualizing the model using diagrams and matrices.

The different classes created in the model are as follow:

Table 1: Ontology classes description

Class name	*characteristics*
Topic	Represent a topic discussed in a Verse
Chapter	Represent a chapter in the Qur'an
Verse	Represent a Verse in a chapter
Word	Represent a term in a verse, a word can be composed of several parts
PronounRef	Represent the relation between a pronoun and its reference. This relation can be one of three types: reference in the same verse, reference in another verse, implicit reference

Define the properties of classes and there facets

We defined the relation between the ontology classes using object properties as described in the following diagram:

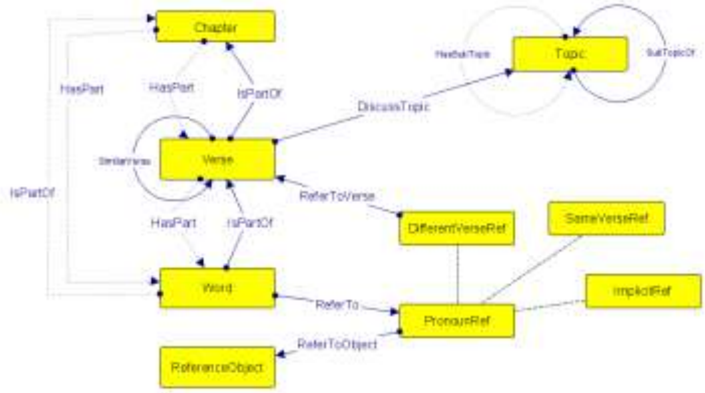


Fig 4: Ontology classes and object properties

The properties using dotted lines are obtained using inference, which means that we won't have a corresponding triple in the data but it will be calculated by the reasoning tool from other triples.

We also defined a number of data properties that will add more information and description of the ontology classes.

Create instances

Creating the instances for our ontology was done by extracting data from the sources described above. These sources were from different formats: OWL, XML and Text. For each extraction we used a program that parse, validate and transform the source data to RDF triples.

The resulting file contains about 1 million RDF triple. Here are some statistics of the obtained RDF triples:

Table 3: Ontology classes statistics

class	*Instances count*
Topic	1181
Chapter	114
Verse	6236
Word	77430
PronounRef	24674

4. EXPLOITING THE RESULTS

Protégé cannot load a big file of RDF triples, so we have to store the ontology in a triple database. There is already a great number of RDF triple store that support SPARQL query language 0(Stegmaier et al., 2009). We will use Jena TDB with Fuseki server for several reasons:

- It has a good performance according to the tests done in 0(Stegmaier et al., 2009).
- The documentation can be found on the project page and is widely complete.
- It provides an ontology API that enables to work on ontologies of different formats, like OWL or RDFS.
- It’s open source.

After loading the data into Jena TDB, we can issue SPARQL queries against the database by using the server user interface or by developing a program using JENA API and the SPARQL ENDPOINT if we want to do more processing with the result. Here are some sample queries that the model can answer:

1) Get the top 10 most discussed topics in the chapters revealed in Mecca

topic	*topicAr*	*Full TopicName*	*Verse count*
Singularity and Being Unique	وحدانيته	أركان الإسلام:التوحيد:توحيد الله تعالي:وحدانيته	405
Promise and Threat	الوعد والوعيد	أركان الإسلام:التوحيد:توحيد الله تعالي:الوعد والوعيد	403
The Characteristics of the Disbelievers	صفات الكفار	أركان الإسلام:التوحيد:الكافرون:صفات الكفار	401
The Quran's Reality and its confirmation of the Previous Books	حقيقته وتصديقه للكتب الأوائل	القرآن الكريم:حقيقته وتصديقه للكتب الأوائل	291
Allah's Address to Him(PBUH)	مخاطبة الله إياه	أركان الإسلام:محمد (ﷺ):مخاطبة الله إياه	228
His Promise to them	وعده إياهم	الإيمان:المؤمنون:وعده إياهم	209
The Companions of the Garden	أصحابها	الإيمان:الغيب:الجنة:أصحابها	203
What has Allah Prepared for them	ما أعده الله لهم	الإيمان:المؤمنون:ما أعده الله لهم	203
Its Fellows	أصحابها	الإيمان:الغيب:النار:أصحابها	180
The Apostates Who Denies the Resurrection	الملحدون المنكرون ليوم البعث	أركان الإسلام:التوحيد:الملحدون المنكرون ليوم البعث	154

2) Get the verses (top 10) that discuss the topic of Zakat without having the string "زكاة" in the verse text

verse Id	*Verse Text*
quran63-10	وأنفقوا من ما رزقناكم من قبل أن يأتي أحدكم الموت فيقول رب لولا أخرتني إلى أجل قريب فأصدق وأكن من الصالحين
quran57-18	إن المصدقين والمصدقات وأقرضوا الله قرضا حسنا يضاعف لهم ولهم أجر كريم
quran2-263	قول معروف ومغفرة خير من صدقة يتبعها أذى والله غني حليم
quran64-16	فاتقوا الله ما استطعتم واسمعوا وأطيعوا وأنفقوا خيرا لأنفسكم ومن يوق شح نفسه فأولئك هم المفلحون
quran64-17	إن تقرضوا الله قرضا حسنا يضاعفه لكم ويغفر لكم والله شكور حليم
quran9-67	المنافقون والمنافقات بعضهم من بعض يأمرون بالمنكر وينهون عن المعروف ويقبضون أيديهم نسوا الله فنسيهم إن المنافقين هم الفاسقون
quran2-274	الذين ينفقون أموالهم بالليل والنهار سرا وعلانية فلهم أجرهم عند ربهم ولا خوف عليهم ولا هم يحزنون
quran9-75	ومنهم من عاهد الله لئن آتانا من فضله لنصدقن ولنكونن من الصالحين
quran22-35	الذين إذا ذكر الله وجلت قلوبهم والصابرين على ما أصابهم والمقيمي الصلاة ومما رزقناهم ينفقون
quran9-104	ألم يعلموا أن الله هو يقبل التوبة عن عباده ويأخذ الصدقات وأن الله هو التواب الرحيم

We can see that we use other words to deal with the topic of "Zakat" like: "أنفقوا", "صدقة", "يقبضون أيديهم", "تقرضوا الله"

3) Get the most used implicit pronoun references

reference	*reference Ar*	*frequency*
Allah	الله	1969
Prophet Muhammad	محمد	1042
mankind	الناس	851
polytheists	المشركين	735
the infidels of Quraish	كفار قريش	689
(Kaafir) the infidels	الكافرين	522
Muslims	المسلمون	520
believers	المؤمنين	485
the hypocrites	المنافقين	353
those who believe	الذين آمنوا	353

5. CONCLUSION AND FUTURE WORK

The present work which deals with the knowledge extraction from the Qur'an showed us that this subject is currently the point of interest of different research groups; but until now there is no global ontology that represents the knowledge contained in the Qur'an. In this work we created a Qur'an ontology that encompasses a set of concepts and the relations between them using OWL. This ontology can be used to answer complex queries and to describe about 11000 resources using over 1 million RDF triple. The next step of our research is to extract more concepts and knowledge from the Qur'an and create a tool that enables users to browse the content of the ontology. As discussed earlier, creating an ontology that covers the content of all the Qur'an would be a complex and time-consuming task and would also require a considerable group of developers and domain experts. We hope that our work will be useful to accomplish this task.

8. REFERENCES

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing, International journal of human-computer studies, 43(5), 907-928.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. Intelligent Systems, IEEE, 21(3), 96-101.

Beseiso, Majdi, Abdul Rahim Ahmad and Roslan Ismail. Article: A Survey of Arabic language Support in Semantic web. International Journal of Computer Applications 9(1):35–40, November 2010. Published By Foundation of Computer Science

Saad, S., Salim, N., & Zainal, H. (2010). Towards context-sensitive domain of Islamic knowledge ontology extraction. International Journal for Infonomics (IJI), 3(1), 197-206.

TA'A, Azman, ABIDIN, Syuhada Zainal, ABDULLAH, Mohd Syazwan, et al. AL-QURAN THEMES CLASSIFICATION USING ONTOLOGY. In: Proceedings of the 4thInternational Conference on Computing and Informatics, ICOCI. 2013. p. 28-30.

Sharef, N.M., Murad, M. A. A., and Mustapha, A., Shishehchi, S., "Semantic Question Answering of Umra Pilgrims to Enable Self-Guided Education", 13th International Conference on Intelligent Systems Design and Applications (ISDA 2013), pp. 141-146, Kuala Lumpur

Sherif, Mohamed Ahmed, Ngonga Ngomo, Axel-Cyrille: Semantic Quran: A Multilingual Resource for Natural-Language Processing. In: Semantic Web Journal (2014), S. 1-5

Dukes, Kais (2013). Statistical Parsing by Machine Learning from a Classical Arabic Treebank. PhD Thesis. University of Leeds.

Aliyu Rufai Yauri, Rabiah Abdul Kadir, Azreen Azman and Masrah Azrifah Azmi Murad, 2013. Quranic Verse Extraction base on Concepts using OWL-DL Ontology. Research Journal of Applied Sciences, Engineering and Technology, 6(23): 4492-4498.

A.M. Sharaf and E. Atwell, "QurAna: Corpus of the Quran annotated with Pronominal Anaphora", in Proc. LREC, 2012a, pp.130-137.

Sharaf, A. B. M., & Atwell, E. (2012b). QurSim: A corpus for evaluation of relatedness in short texts. In LREC (pp. 2295-2302).

Abbas, N. H. (2009). Quran Search for a Concept Tool and Website. PhD Thesis. University of Leeds.

Psyché, Valéry, Olavo Mendes, and Jacqueline Bourdeau. "Apport de l'ingénierie ontologique aux environnements de formation à distance." Revue des Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF) 10 (2003): 89-126.

NOY, N. F. (2001). Ontology Development 101: A Guide to Creating Your First Ontology: Knowldege Systems Laboratory, Stanford University. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Pinto, H. S., & Martins, J. P. (2000). Reusing ontologies. In AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes (Vol. 2, No. 000, p. 7). AAAI Press.

Stegmaier, Florian, et al. "Evaluation of current RDF database solutions." Proceedings of the 10th International Workshop on Semantic Multimedia Database Technologies (SeMuDaTe), 4th International Conference on Semantics and Digital Media Technologies (SAMT). 2009.