An Ontological Approach to Intelligence Gathering using Semantic Metadata

Tameem Chowdhury, Dr. Stilianos Vidalis

October 10, 2012

Abstract

We discuss our research into combatting the problem of effectively presenting next to real-time knowledgeable answers to runtime user generated queries, from disparate sources. The paper explores the foundations of our research for the design of an Information Gathering tool based on the Intelligence Domain; focusing on the exploration of semantic metadata, incorporating ontologies and implementing swarm intelligence theory in the conceptualisation of the system design for the IGUSMON Project, in order to present an efficient and innovative solution.

1. Introduction

Through the implementation of consistent semantic metadata and well defined ontologies for different entities, assets within the business context and information within the consumer sphere can be efficiently stored and organised; the result of which ensures easily retrievable information, data, resources and assets. "The World Wide Web was originally built for human consumption, and although everything on it is machine-readable, this data is not machine-understandable. It is very hard to automate anything on the web, and because of the volume of information the web contains, it is not possible to manage it manually." [1]

For the successful development of an Intelligence Gathering tool that will collect, structure and present the data in the form of knowledgable answers, ontologies, based on the intelligence domain, will be formulated and this will incorporate semantic metadata in the collection and structure process. The aim of the system will be to provide next to real-time knowledgeable answers to runtime user generated queries, from disparate sources, in noncritical multimedia systems, henceforth referred to as the IGUSMON Project. We present our design, which combines ideas discussed in "The Semantic Web" [2] with theory proposed from the study of nature, most notably for our research, Swarm Intelligence [3], will be implemented into the IGUSMON Project design.

The outline for the paper is as follows: Section two will discuss Metadata and

Semantics and the advantages of having well defined concepts for the appropriation of semantic asset metadata; Section three provides insight into ontology and the benefits of its implementation; Section four explores Swarm Intelligence [3] and how the theory studied and documented from research into particular natural systems can help design an efficient computer system with the ability to utilise logic in its decision making and Section five will conclude with the conceptualisation of the proposed system for the IGUSMON Project.

2. Metadata and Semantics

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information. [4]

Carrier (2005) also concurs; the metadata category contains the data that describe a file; they are data that describe data.[5] In its most generalised form, metadata include information such as where file content is stored, how big is the file, the times and dates when the file was last read from or written to and the access control information. In terms of file attributes, the metadata does not relate to the file name or the content of a particular file. Metadata is utilised in the classification, archiving and most importantly the retrieval of information, data, resources and assets. If the metadata is maintained and organised correctly, the availability and retrieval is exponentially increased.

"The three main types of metadata:
1. Descriptive Metadata describes a resource for purposes such as discovery and identification.
2. Structural Metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
3. Administrative Metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information and who can access it."[4]

Semantic is defined as the branch of linguistics and logic concerned with meaning. The two main areas are logical semantics, concerned with matters such as sense, reference, presupposition and implication, and lexical semantics, concerned with the analysis of word meanings and relations between them.[6] Semantic Metadata, or meaningful and useful data, are essential in todays information oriented world of discovery and will be the basis of developing ontologies for the intelligence domain, essential for the efficient retrieval, organisation and interpretation of data.

"The Semantic Web", [2][7] is still sporadic in global application of web design and development, but the theory proposed is very applicable for our purposes. By creating semantic metadata for all data, the archival and most importantly retrieval is supplemented and accelerated.

Metadata is utilised in a variety of different situations by varying institutions. The Police Force, Military facilities, Governments, Libraries, Museums, Internet

search engines, Public and Private Sector companies are just a few examples of where metadata is applied and incorporated into everyday tasks and utilised on a daily basis. Foulonneau et Riley [8] add: "Metadata allows various functions to be performed on digital resources, for example, discovery, interpretation, preservation, management, representation and the reuse of objects."
Metadata is found everywhere. Objects, Data and Assets are synonymously identified through their metadata and in terms of digital resources, enable many functions to be conducted utilising them. Table 1 lists the different metadata standards and schema that are applied to the organisation and archival of different resources in different fields, which enable the retrieval, interpretation and reuse of a particular resource.

As in many fields and industries, standards are employed; ensuring compatibility, interoperability and repeatability and this applies to metadata and its application.There are already many different metadata standards and schema that exist in order to provide a global initiative in organising different information, data, resources and assets. "Many of these initiatives are based on or are compatible with the ISO Reference Model for an Open Archival Information System (OAIS)."[4]
The schema provide an array of elements that can be applied within the conceptual design of an ontology and for the purposes of our research, each of the schemas will be analysed in order to determine whether to use elements or create a new hybrid and more specific framework for creating ontologies for the intelligence domain.

Table 1: Metadata Schema

Metadata Schema Classifications
1. MARC 21
2. AACR2
3. The Dublin Core Metadata Element Set
4. The Text Encoding Initiative (TEI)
5. Learning Object Metadata (LOM)
6. The Interoperability of Data in E-Commerce Systems Framework (indecs)
7. Online Information Exchange International Standard (ONIX)
8. Categories for the Descriptions of Work of Art (CDWA)
9. The VRA Core Categories Element Set
10. The ISO/IEC Moving Picture Experts Group (MPEG) Multimedia Metadata
11. MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938)
12. MPEG-21, Multimedia Framework (ISO/IEC 21000)
13. Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)
14. Data Documentation Initiative (DDI)
15. Z39.50 Protocol.

3. Ontologies

Simply defining ontology is exigent and requires some background into its lexicology and etymology. .

Ontology, in the original sense, relates to metaphysics and the study of existence and being in the most philosophical sense. Logically, ontology can be defined as "the set of entities presupposed by a theory."[9]

Ontology is a "systematic account of existence; An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them." [10]

Jokela (2001) concurs:"Ontologies are conceptual models that map the content domain into a limited set of meaningful concepts." [11] Formal ontology aims to provide a specification of the meaning of terms within a vocabulary. When conceptualising ontological expressions, the design needs to ensure that the continuants and participants are not stochastically determined. [12] Ontologies must contain classes and use relations that are based on Aristotelian universals. Aristotelian realism states that universals and their instances share a symbiotic relationship; one cannot exist without the other. In particular, there are no universals, which have no instance in reality. Dumontier et Hoehndorf [12] describe a Basic Formal Ontology (Figure 1) which will assist in the initial creation of entities and their continuants.

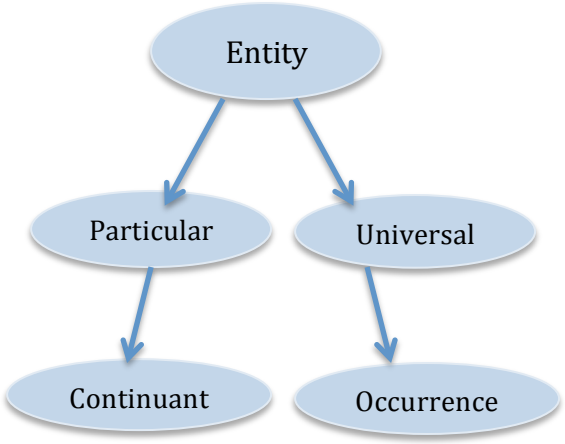

Figure 1: Basic Formal Ontology [12]

By defining an ontology based on a particular domain, and in the case of our research, intelligence gathering, the tool can be designed return information in a structured manner and only the information defined within the ontology. Oldfield [13] states, "A domain model is a model of the domain within which an Enterprise conducts its business."

Domain models enable the efficient organisation and management of a businesses assets and the Domain Model for one Enterprise should be the same as that for any other Enterprise conducting business in the same domain. [13] Domain models can be thought of as a domain of interest, which describes the various entities, their attributes, roles and relationships, as well as the constraints that govern the integrity of the model elements, or instances when referring to the ontology classification. The domain model is created in order to represent the vocabulary and key concepts of the specified domain.[14]

Ontology enables a computer to take as input predefined statements or rules and use them to simulate logical decision making. Ontologies are very extensible and application independent, enabling future improvement and growth at a level that will not impact the application of the system. Mankato [15] presents an excellent breakdown of the thinking required behind

the ontology design, illustrated in figure 2.

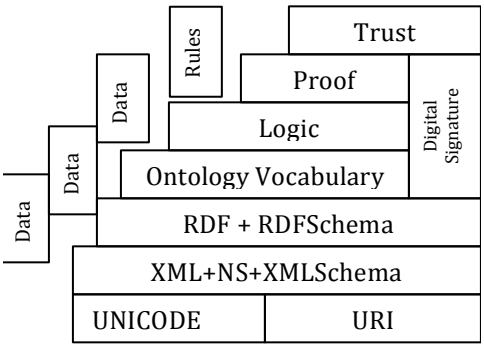

Figure 2: Visualisation of layers for Ontology Creation [15]

A Common machine readable language is needed to to implement ontologies, and a number of different languages are listed in Table 2.
RDF recommends and implements a function of creating 'Triplets' of statements which are the stored and cross referenced by the system in its logical decision making process. A 'Triplet' consists of a *Subject*, *Verb/Predicate* and an *Object*.[1] This is illustrated in Figure 3.

Table 2: Ontology Languages

Languages for Creating Ontologies
1. Resource Development Framework (RDF)
2. DARPA Agent Markup Language + Ontology Inference Layer (DAML + OIL)
3. Ontology Web Language (OWL)
4. OWL Lite

By defining and implementing these triplets to form interrelationships, effective ontologies will be created for the IGUSMON Project. This is the inception of our solution to the problem statement of our research, where the combination of utilising the semantic metadata with the creation of ontologies focusing on the Intelligence domain, will enable the system to implement logic in its decision making. The notions put forth by Dumontier et Hoehndorf [12] will also be considered ensuring that *Entities* or *Subjects* can be combined with meaningful *Continuants* or *Objects*' respectively.

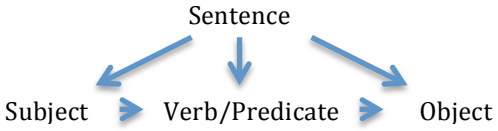

Figure 3: RDF 'Triplet' elements [1]

4. Swarm Intelligence

"Theories of self organisation (SO) [16][17] originally developed context of physics and chemistry to describe the emergence of macroscopic patterns out of processes and interactions defined at the macroscopic level, can be extended to social insects to show that complex collective behaviour may emerge from interactions along individuals that exhibit simple behaviour: in these cases, there is no need to invoke individual complexity to explain complex collective behaviour. Recent research shows that SO is indeed a major component of a wide range of collective phenomena in social insects [18]."

"Social insects have limited cognitive abilities: it is, therefore, simple to design agents, including robotic agents, that mimic

their behaviour at some level of description."Swarm Intelligence, theories developed through research and study into natural systems, are implemented and utilised in design elements of robotic agents, are ideal for the IGUSMON Project.

An important note; the systems of nature discussed here and their behaviours are theories, in the continuous processes of study and research and the accuracy of the exact biological science of their physical behaviour is not of importance for our purposes, nor is it unequivocable fact; as Bonebeau et al. [3] quite acutely comment: "algorithms do not have to be designed after accurate or true models of biological systems;: efficiency, robustness and flexibility are the driving criteria, not biological accuracy."
The modelling of social insects by means of SO can help design artificial distributed problem solving devices that self organise to solve problems - swarm-intelligent systems. It is, however, fair to say that very few applications of swarm intelligence have been developed. One of the main reasons for this relative lack of success resides in the fact that swarm-intelligent systems are hard to "program", because the paths to problem solving are not predefined but emergent in these systems and result from interactions among individuals and between individuals and their environment as much as from the behaviours of the individuals themselves. Therefore, using a swarm-intelligent system to solve a problem requires a thorough knowledge not only of what individuals behaviours must be implemented but also of what interactions are needed to pro due such or such global behaviour. This is where we propose to introduce ontology in the design of the system."

"In a social insect colony, a worker usually does not perform all the tasks, but rather specialises in a set of tasks, according to its morphology, age or chance. This division of labour among nest mates, whereby different activities are performed simultaneously by groups of specialised individuals, is believed to be more efficient than if task were performed sequentially by unspecialised individuals. [19][20]" Aspects of Swarm Intelligence, specifically the behaviours of worker insect colony ants, will be implemented into the design of the ontologies and the operation of the different spiders (Section five) the system will utilise, in the collection and structure of data, requested from runtime user generated queries.

Jeanne [21] provides another example:"Nest construction in the wasp *polybia occidentals* involves three groups of workers, pulp foragers, water foragers and builders. The size of each group is regulated according to colony needs through some flow of information among them [21]"

5. Conceptualisation of Intelligence Tool Architecture

Web spiders enable the search and retrieval of specific information from the contents of a particular webpage or website. Furthermore spiders can be programmed to search vast datasets without the need for continuous human interaction. Once the spider is deployed it can crawl from webpage to webpage, through the extraction of hyperlinks and therefore create a list of searchable content for the web spider.

They can be implemented into Intelligence Gathering as they are programmed to collect defined and specific information from disparate sources, relationally stochastic and orthogonal, providing a required independancy. By examining the metadata of the digital resource and therefore they are a compulsory component of the IGUSMON Project design. The web spiders provides an excellent mechanism for gathering the required websites and the corresponding semantic metadata for the target search, which will then enable the other features of the system to mine and structure the data for presentation in the form of a knowledgable answer.

Figure 4 lustrates the conceptual design for the tool, which will incorporate web spiders as a mechanism for gathering the raw data before a validation module, incorporating the specific ontology and a data-mining algorithm will analyse and structure the data into information.

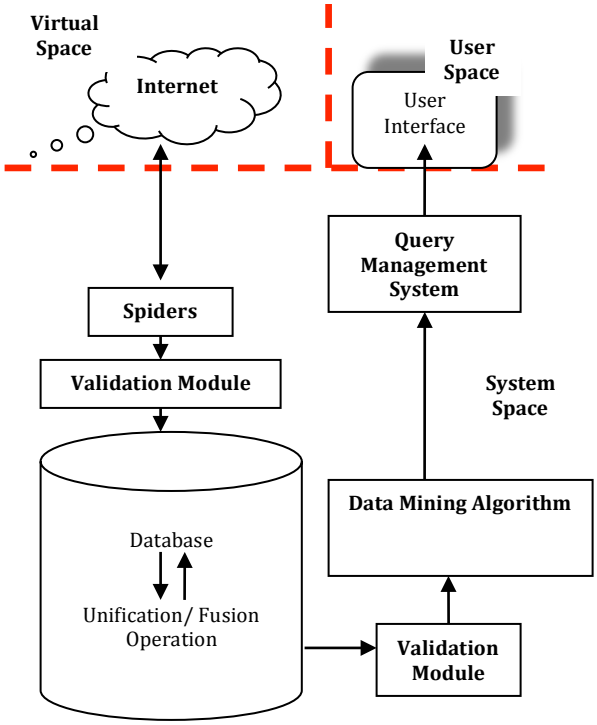

Figure 4: IGUSMON Project System Architecture

Data Mining [22] (sometimes called data or knowledge discovery) is a process of analysing data from different perspectives and summarising it into useful information information that can be used to increase revenue, cut costs or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions, angles, categorise it and summarise the relationships identified. Technically data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [22] , and the concept of analysing collated data will apply for the Intelligence Gathering Tool and it will do so based on the meta tags of webpages and websites.

Blankson [23] states that, "Meta Tags tell search engines how to handle, index and display pages." [23] Meta tags are special Hyper Text Mark-up Language (HTML) tags that provide information about a webpage. Unlike normal HTML tags, Meta tags do not affect how the page is displayed. Instead they provide information such as who created the page, how often it is updated, what the page is about and keywords regarding the page content. Many search engines use this information when building their indices.

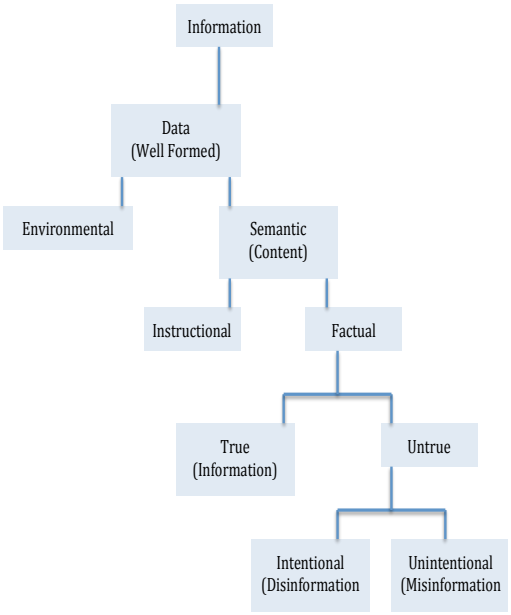

Figure 5: Mathematical Theory of Communication [24]

The meta tags present within webpages and websites will provide much of the semantic metadata that the tool will analyse and extract only the information defined within the proposed ontology. The information gathered will be filtered through a data-mining algorithm, and the architecture of the validation will incorporate Floridis [24] Mathematical Theory of Communication (MTC) in the design, illustrated in Figure 5.

6. Conclusions

Incorporating the theories of Berners-Lee et al.[2], Dumontier and Hoehndorf [12], And Bonabeau et al. [3] will enable effective ontologies to be created for the Intelligence domain which can be utilised in the IGUSMON Project, that will provide knowledgable answers to user generated runtime queries. The semantic metadata of digital resources will be located and utilised in the formation of defining triplets, which which provide the proposed system to simulate the appearance of logic in its decision making in terms of intelligent data collection and structure.

To be added too...

References

[1] W3C.2012.[WWW] http://www.w3.org/TR/1998/WD-rdf-syntax-19981008/ (28th August,2012)

[2] BERNERS-LEE, T., HENDLER, J. ET LASSILA, O. 2001. *The Semantic Web.* Scientific American Feature Article.

[3] BONABEAU, E., DORIGO., M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems.* New York: Oxford University Press.

[4] NATIONAL INFORMATION STANDARDS ORGANISATION PRESS. 2004. *Understanding Metadata.* Bethesda: NISO Press.

[5] CARRIER, B. 2005. *File System Analysis.* New Jersey: Pearson Education, Inc.

[6] OXFORD UNIVERSITY PRESS. 2012. [WWW] http://oxforddictionaries.com/definition/semantics (23rd January, 2012)

[7] CARDOSO, J. 2010. *The Semantic Web: A Mythical story or a solid reality?* Metadata and Semantics. New York: Springer Science+Business Media.

[8] FOULONNEAU, M. ET RILEY, J.2008. *Metadata for Digital Resources*. Oxford: Chandos Publishing (Oxford) Limited.

[9] COLLINS ENGLISH DICTIONARY. 2012. [WWW]. http://dictionary.reference.com/browse/ontology. 10th Edn. HarperCollins Publishers. (13th January 2012)

[10] HOWE, D. 2010. [WWW] http://dictionary.reference.com/browse/ontology (23rd January, 2012)

[11] JOKELA, S. 2001. *Metadata Enhanced Content Management In Media Companies*. Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 114, Espoo: Finnish Academies of Technology.

[12] DUMONTIER, M. ET HOEHNDORF, R. 2010. *Realism for Scientific Ontologies*.6th International Conference on Formal Ontology in Information Systems. P.387-399. University Of Liepzig.

[13] OLDFIELD. 2002.

[14] MARTIN, B. ET MITROVIC, A. *ITS Domain Modelling with Ontology: Example of a Useable Ontology*.

[15] MANKATO, P. 2011. [WWW] *The Semantic Web - An Overview*. http://www.youtube.com (10th September 2012)

[16] HAKEN, H. 1983. *Synergetics*. Berlin: Springer-Verlag. *Cited in* BONABEAU, E., DORIGO., M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems*. New York: Oxford University Press.

[17] NICOLIS, G., ET PRIGOGINE, I. 1977. *Self-Organiization in Non-Equilibrium Systems*. NewYork, NY: Wiley & Sons. *Cited in* BONABEAU, E., DORIGO., M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems*. New York: Oxford University Press.

[18] DENEUBOURG, J., GOSS, S., FRANKS, N. ET PASTEELS, J. 1989. *The Blind Leading the Blind: Modelling Chemically Mediated Army Ant raid Patterns*. Insect Behav.2: 719-725. *Cited in* BONABEAU, E., DORIGO., M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems*. New York: Oxford University Press.

[19] JEANNE, R. 1986. *The Evolution of the Organization of Work in Social Insects*. Monit.Zool.Ital. 20: 119-133.*Cited in* BONABEAU, E., DORIGO., M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems*. New York: Oxford University Press.

[20] ROBINSON, G. 1992. *Regulation of Vision of Labour In Insect Societies*. Annu.Rev. Entomol. 37: 637-665. *Cited in* BONABEAU, E., DORIGO., M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems*. New York: Oxford University Press.

[21] JEANNE, R. 1996. *Regulation of Nest Construction Behaviour in Polybia occidentalis*. Anim. Behav.52: 473-488. *Cited in* BONABEAU, E., DORIGO.,

M ET THERAULAZ, G. 1999. *Swarm Intelligence From Natural to Artificial Systems*. New York: Oxford University Press.

[22] PALACE, B. 1996. Data Mining

[23] BLANKSON, S. 2007. Meta Tags: Optimising Your Website for Internet Search Engines. London: Blankson Enterprises Limited.

[24] FLORIDI, L. 2011.