Bridging Semantics Through Ontologies

Tameem Chowdhury
School of Design, Engineering and Technology
University of South Wales
Newport, Wales.
Email: tameem.chowdhury@southwales.ac.uk

Christopher Tubb
School of Design, Engineering and Technology
University of South Wales
Newport, Wales.
Email: Christopher.tubb@southwales.ac.uk

Stilianos Vidalis
Faculty of Computing, Engineering and Technology
Stafford University
Stafford, England.
Email: Stilianos.vidalis@staffs.ac.uk

***Abstract*— Semantic metadata enables contextual and relevant data to be identified for a particular entity. The use of ontologies creates a bridging mechanism, whereby semantic metadata can be referenced and validated to ensure that relevant and useful information is collected. This also ensures trust and logic can be attained in search functionality. The paper explores the foundations of the research for the design of an Information Gathering tool for the Business Intelligence Domain. The aim of the project is to effectively present next to real-time knowledgeable answers to runtime user generated queries for extracting business intelligence. The tool will collect information from disparate sources and requires the implementation of semantics to safeguard the future of knowledge discovery and reuse. This paper summaries the research and conceptualisation for our Information Gathering tool using semantic metadata to be utilised in the area of Business Intelligence.**

Keywords-Semantics; Metadata; Ontology; Business Intelligence.

I. INTRODUCTION

"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].

As the technological growth exponentially increases, the vastness of data and information available for consumption and reuse is equally daunting. Incorporating semantics, specifically semantic metadata, into search functionality and classification, relevance and precision can be enhanced. In order to successfully implement semantic metadata, ontologies can be utilised and these principles can be applied for conducting knowledge extraction for gaining Business Intelligence (BI). The paper discusses the fundamentals of semantic metadata and ontology and how their application will benefit the Intelligence Gathering Using Semantic Metadata and Ontology (IGUSMON) project, currently work in progress. The aim of the tool is to provide next to real-time knowledgeable answers to runtime user generated queries, from disparate sources, in noncritical multimedia systems focusing on BI. We present the 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] and proposes how they can be applied to extract knowledge for BI.

The outline for the paper is as follows: Section II will discuss the fundamentals of semantic metadata and the advantages of having well defined concepts for appropriation. It further explores Swarm Intelligence 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. Section III presents the design of the IGUSMON project algorithm and analyses the benefits and limitations that may be encountered during the development phase. Related and existing work is also identified.

II. SEMANTIC METADATA AND ONTOLGICAL FUNDAMENTALS

For the design of an Intelligence Gathering tool, the difference between simple information, assets and actual intelligence required definition and identification. Information encapsulates a wide range of concepts and phenomena. They relate to both the processes and material states, which are closely interrelated. Information can be:

- "A product, which encompasses information as an object, as resource, as commodity.

 ISBN: 978-1-61208-293-6

- What is carried in a channel, including the medium channel itself.
- The Contents." [4][5][6][7]

Information can be an asset to stakeholders and or a particular entity, for example, data companies such as Acxiom, TargusInfo and BlueKai [8]. An asset can be defined as a single item of ownership having exchange value [9][10][11][12][13]. Information assets are physical, hardware, software, data, communications, administrative and personnel resources of a computing system [14]. Every information asset contains some sort of information that we can analyse and extract intelligence from.

Intelligence can be defined as a specialised form of knowledge, an activity, and an organisation. As knowledge, intelligence informs leaders, stakeholders or entities, uniquely aiding their judgment and decision-making. As an activity, it is the means by which data and information are collected, their relevance to an issue established, interpreted to determine likely outcomes, and disseminated to individuals and organisations who can make use of it, otherwise known as consumers of intelligence [15]. This becomes more complicated depending on the situation and the stimuli that we are observing and impacts how we extract different intelligence. The application and usability of this intelligence simply depends upon the search criteria and purpose for the collection. For the objectives of the IGUSMON project, collected information will be referenced against ontologies, which will be specifically created for BI, to filter relevant intelligence according to the subjects identified.

An important factor when collecting information that will be classified, as intelligence is the need for accuracy and trust, since the World Wide Web or information environment, unfortunately and inevitably provides a wealth of misinformation. The United States Department of Defense (DoD) has defined the Information Environment (IE) as:

"The aggregate of individuals, organisations and systems (resources) that collect, process, disseminate, or act on information." [16]

Akin to reality, the virtual space is the new realm of warfare and dissemination of misinformation. Clausewitz and Tzu [17][18] theorised about warfare and military mentality and strategy in their respective works, and although the context is different, the theory can still be applied to virtual information warfare. Through the implementation of consistent semantic metadata and well-defined ontologies, BI will be collected, structured, efficiently stored and organised; ensuring they can also be easily retrieved and analysed when required. Furthermore the threat of misinformation can be minimised and or eliminated and trust attributed to the extracted knowledge. 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 [19]. It is utilised in the classification, archiving and most importantly the retrieval of information, data, and resources and assets. If the metadata is maintained and organised correctly, the availability and retrieval is exponentially increased [20][21].

Jokela [20] identifies thirteen categorisations of metadata, of which we have identified the three main types of metadata that will be utilised in the IGUSMON project:

- Descriptive Metadata describes a resource for purposes such as discovery and identification.
- Structural Metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- 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 [19][20][21].

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 [22]. Foulonneau and Riley [21] add: "Metadata allows various functions to be performed on digital resources, for example, discovery, interpretation, preservation, management, representation and the reuse of objects."

Semantics is 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 [23]. Semantic Metadata, or meaningful and useful data, are essential in today's information oriented world of discovery and provide the foundations for developing our ontologies.

Simply defining ontology is exigent and requires some background into its lexicology and etymology. Originally the term is from philosophy and denotes a systematic account of existence. In computer science and Artificial Intelligence, ontology is an explicit specification of a conceptualisation and states what exists can be represented [24].

Jokela [20] concurs: "Ontologies are conceptual models that map the content domain into a limited set of meaningful concepts." Formal ontology aims to provide a

 ISBN: 978-1-61208-293-6

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 [25].

By defining ontologies based on a particular domain [26], the algorithm [27][28] within the Intelligence Gathering tool will facilitate the return of intelligence in a structured manner and only for information predefined within our ontologies for BI. Figure 1 presents a breakdown of the thinking required behind ontology design and will form the foundations for developing our BI ontologies.

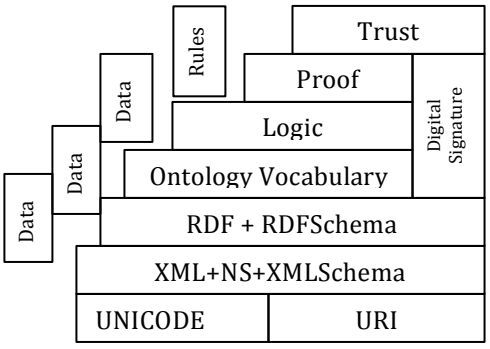

Fig. 1. Visualisation of layers for Ontology Creation [2][29]

The combination of utilising the semantic metadata with the creation of ontologies focusing on the intelligence domain, integrated by an algorithm, will enable the system to simulate and implement logic in its decision-making. The notions put forth by Dumontier and Hoehndorf [25] will also be considered, ensuring that Entities or Subjects can be combined with meaningful Continuants or Objects' respectively [22]. The algorithm will utilise web spiders to collect the data, and use swarming agents to enable communication between the different system components, which include the ontologies.

Swarm intelligence [3][30] theories, developed through research and study into natural systems, are often implemented and utilised in the design of robotic agents. "Theories of Self-Organisation (SO), were originally developed for the contextual benefit of physicists and chemists to describe the emergence of macroscopic patterns" [31][32]. However, SO can be extended to social insects and describe how complex collective behaviour may emerge from interactions along individuals that exhibit simple behaviour but contribute towards the same task. Recent research reflects that SO is indeed a major component of a wide range of collective phenomena in social insects and designers of robotic agents have applied this natural inspiration in the realisation of different robotic agents and artificial systems [22][33]. Social insects have limited cognitive abilities, and therefore the simplicity can be applied to the design of robotic agents, that mimic their behaviour at some level of description [3][31].

The systems of nature 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. "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" [3]. This is why we often use the term biologically inspired. The modelling of social insects by means of SO can help design artificial distributed problem solving devices- swarm-intelligent systems. Although biologically inspired swarm intelligence has an appeal to those developing such 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, resulting from interactions among individuals and between individuals and their environment, as much as from the behaviours of the individuals themselves [3]. There are two types of emergence, light and strong. Light emergence, where the final behaviour can be deduced from the rules, is in contrast to strong emergence. There are philosophical arguments regarding this; however it is always easier to take a system and analyse how the behaviour results from the interacting rules, than it is in all but trivial cases, to engineer behaviour from simple interacting rules. Therefore, using a swarm-intelligent system to solve a problem requires a thorough knowledge not only of what individual's behaviours must be implemented but also of what interactions are needed to produce such or such global behaviour [3]. This is where ontologies are introduced into the design of our system.

The reduction of the behaviour of these agents can be expressed in equations [3] and have been applied in applications in the areas of Robotics, Information Operations, Evolutionary Computing, Neural Networks, Agent Management and others [30]. Watson adds, "Agent properties can be utilised in: Learning; Social Learning; Environmental Learning; Histories; Cognition and Communications" [30].

III. IMPLEMENTING ONTOLOGIES WITH SEMANTIC METADATA WITHIN THE IGUSMON PROJECT

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 hyper-

 ISBN: 978-1-61208-293-6

links and therefore create a list of searchable content. Spiders can implement intelligence gathering through the collection of specific information from disparate sources, relationally stochastic and orthogonal. They can be programmed for the required level of independency, and will function by examining the semantic metadata of the digital resource. The web spiders provide 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 knowledgeable answer [22].

The research is in its infancy and the following architecture and design described is the conceptualisation of our algorithm for intelligence gathering using semantic metadata. Figure 2 illustrates the conceptual design of the Information Gathering tool, which demonstrates how the web spiders will act as a mechanism for gathering the raw data, before sending the extracted semantic metadata back to the database for validation with the predefined ontologies. Once the extracted data is verified, a data-mining [34] algorithm structures the data into information before returning it as a knowledgeable answer to the Query Management System.

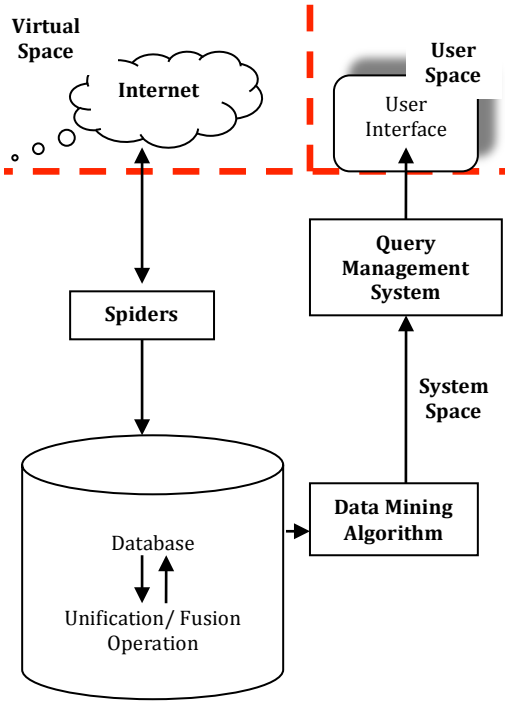

Fig. 2. IGUSMON Project System Architecture [22]

Venturing deeper into the mechanics of the Intelligence Gathering tool and specifically to the core elements of the design, Figure 3 illustrates the System Architecture and the critical elements of the system, as well as how the swarming agents communicate. The Query Management System will signal the release of the web spiders from the spider deployment module via the database and a swarming agent. Collected information will be verified for relevant intelligence within the Validation Module via ontology checks. However, before the semantic metadata reaches the Validation Module, a final check will be conducted via a worker agent against the Irrelevant Data module, where discarded information from previous extractions, that did not produce positive intelligence results relating to a query, are stored.

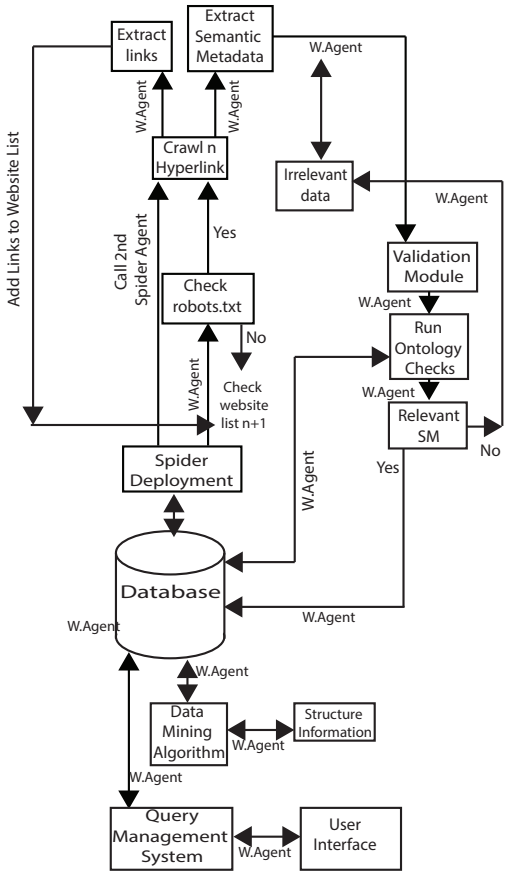

Fig. 3. Implementation of Swarming Intelligence Behaviour with the IGUSMON Project System Architecture

As stated earlier, all information may prove to be intelligence and can be utilised depending upon a particular objective or query; therefore all extracted semantic metadata will be stored, either in the database or the Irrelevant Data module. The information gathered will be filtered through a data- mining [34] algorithm and the architecture of the Data Mining Algorithm will incorporate Floridi's [5] Mathematical Theory of Communication (MTC) in the design, illustrated in Figure 4.

The architecture of the algorithm differs from related work in that it focuses only on extracting semantic meta-data for filtering against our BI ontologies. Furthermore, the application of swarming worker agents within the system ensures that multiple tasks are conducted concurrently. The

 ISBN: 978-1-61208-293-6

benefits of this focus are anticipated to ensure vast datasets can be quickly referenced and utilised for extraction. The direct integration of the semantic metadata with the ontologies will ensure that relevant knowledge can be extracted. An obvious limitation to this method will be determined by how much of the relevant data is attributed with semantic metadata. Even though semantic web methods have been proposed for over a decade now, data does exist that was created before and after, which seldom or minimally focuses on semantics. However this does not mean that semantic metadata is limited; with the technological growth and vast amounts of growing data, this limitation is becoming less finite. The other limitation that will impact our research will be the reach of the algorithm. When the conceptualisation of the algorithm is developed, the testing will focus on a finite number of websites for extraction, due to available computing power and time constraints. As mentioned, the focus of the IGUSMON project is currently immersed within this area and development is in progression; some of the design elements proposed may change as the modules are created and tested for feasibility.

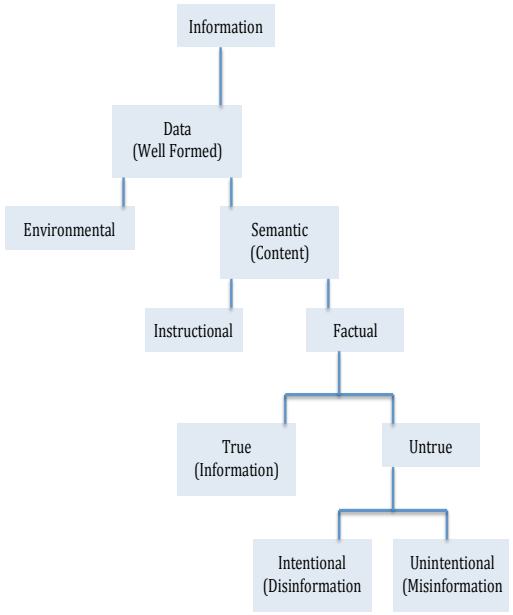

Fig. 4. Mathematical Theory of Communication [5][22]

A. Related Work

The foundations of the research are attributed to Berners-Lee et als. *"The Semantic Web"* [2], Gruber's [24] research on ontologies and Bonabeau et als. [3] research on Swarm Intelligence. Further related work within our research focused on where semantic metadata and ontological mapping has been incorporated within the design, collection and extraction. Jokela [20] implements the use of semantic metadata within media content. Whereas, Stefanov and Huang's [35] research focuses on metadata context management. Vlachidis et al. [36] attribute and refer to this concept of utilising semantic metadata as Semantic Annotation within their research. They incorporate Semantic Annotation within their mechanism responsible for connecting natural language and formal conceptual structures, observing that the incorporation of semantic metadata could enable new information accessibility and enhance existing methods and systems. The IGUSMON project focuses on applying these methods in the area of BI.

IV. CONCLUSIONS AND FUTURE WORK

Implementing ontologies into the design and application of the IGUSMON project, enables relevant information to be defined within a strict set of requirements, so that precise retrieval can be achieved. The sheer volume of information assets or intelligence that can be gathered through search today is overwhelming; the focus on semantic metadata ensures that ontologies can be developed to conceptualise subjects and objects and ultimately enable us to simulate logic in the search for valuable intelligence. The development of the algorithm and the creation of the ontologies for BI have begun. The intention for demonstrating the successful completion of the algorithm and architecture will be through the use of a user interface, enabling users to submit runtime generated queries. The design of the algorithm and overall architecture of the tool, will ensure that if the ontologies are modified, there will be minimal disruption and ensures that any expansion of search parameters can be integrated. Semantics enable contextual and relevant intelligence to be gathered; the extensibility of the database storing the ontologies ensures that additional information and specifically triplets, can be incorporated when a limitation is identified. This is a key factor since the web spiders will retrieve information specified by their defined semantic metadata, and as linguistics and modern languages have taught us throughout history, the semantics of words and expressions are always evolving to reflect changes in society.

ACKNOWLEDGMENT

Tameem Chowdhury would like to acknowledge the European Social Fund for the funding provided through The Knowledge Economy Social Skills (KESS) Ph.D. Scholarship Scheme, enabling the opportunity to conduct this research and development for the IGUSMON Project.

REFERENCES

[1] W3C. RDF Syntax. [www] http://www.w3.org/TR/PR-rdf-syntax/, 1999. [retrieved: July 2013].
[2] T. Berners-Lee, J. Hendler, and O. Lassila. "The semantic web." Scientific American, 284(5): pp. 34–43, 2001.
[3] E. Bonabeau, M. Dorigo, and G. Theraulaz. "Swarm Intelligence From Natural to Artificial Systems." New York: Oxford University Press, 1999.
[4] M. Menou. "The impact of information ii: Concepts of information and its value." Information Processing & Management, 31(4): pp. 479–490, 1995.
[5] L. Floridi. "Semantic conceptions of information." In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Spring 2011 edition, 2011.

 ISBN: 978-1-61208-293-6

[6] N. Belkin. “Information concepts for information science.” Journal of Documentation, 34(1): pp. 55-85, 1978.
[7] B. Brookes. “The foundations of information science. Part i. philosophical aspects.” Journal of Information Science, 2(3-4): pp. 125–133, 1980.
[8] E. Pariser. The Filter Bubble. Penguin Group, 2011.
[9] K. Choong. “Intellectual capital: definitions, categorization and reporting models.” Journal of Intellectual Capital, 9(4): pp. 609–638, 2008.
[10] P. Kostagiolas and S. Asonitis. “Intangible assets for academic libraries: Definitions, categorization and an exploration of management issues.” Library Management, 30(6/7): pp. 419–429, 2009.
[11] L. Joia. “Measuring intangible corporate assets: Linking business strategy with intellectual capital.” Journal of Intellectual Capital, 1(1): pp. 68–84, 2000.
[12] L. Kaufmann and Y. Schneider. “Intangibles: A synthesis of current research.” Journal of Intellectual Capital, 5(3):366–388, 2004.
[13] M. Harvey and R. Lusch. “Protecting the core competencies of a company: Intangible asset security.” European Management Journal, 15(4): pp. 370–380, 1997.
[14] T. Chowdhury, S. Vidalis, and C. Tubb. “Proactively defending computing infrastructures through the implementation of live forensic capture in corporate network security.” In The 3rd International Conference on Cybercrime, Security and Digital Forensics. 2013. In press.
[15] N. Hendrickson. “Critical thinking in intelligence analysis.” International Journal of Intelligence and Counter Intelligence, 21(4): pp. 679–693, 2008.
[16] S. Vidalis and O. Angelopoulou. “Deception and manoeuvre warfare utilising cloud resources.” In The 3rd International Conference on Cybercrime, Security and Digital Forensics. 2013. In press.
[17] C. Von Clausewitz. On War. Princeton: Princeton University Press. 1976.
[18] S. Tzu. The Art of War by Sun Tzu Special Edition (Translated and annotated by Lionel Giles). El Paso Norte Press, 1910.
[19] NISO. Understanding Metadata. National Information Standards Organization: Bethesda, 2004.
[20] S. Jokela. Metadata Enhanced Content Management In Media Companies. Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 114, Espoo: Finnish Academies of Technology, 2001.
[21] M. Foulonneau and J. Riley. Metadata for Digital Resources. Oxford: Chandos Publishing (Oxford) Limited, 2008.
[22] T. Chowdhury, S. Vidalis, and C. Tubb. “An ontological approach to intelligence gathering using semantic metadata.” The Journal of Communication and Computer US. 2013. In press.
[23] Oxford University Press. [www] http://oxforddictionaries.com/definition /semantics, 2012. [retrieved: January 2012].
[24] T. Gruber. “A translation approach to portable ontology specifications.” Knowledge Creation Diffusion Utilization.” 5(April): pp.199–220, 1993.
[25] M. Dumontier and R. Hoehndorf. “Realism for scientific ontologies.” *6th International Conference on Formal Ontology in Information Systems*. volume 209, pp. 387–399. IOS Press, 2010.
[26] M. Oldfield. Domain Modelling. [www] http://www.aptprocess.com, 2002. [Retrieved: March 2012].
[27] S. Dasgupta, C. Papadimitriou and U. Vazirani. Algorithms. McGraw-Hill, 2006.
[28] T. Cormen, C. Leiserson, R. Rivest and C. Stein. Introduction to Algorithms. The MIT Press: Cambridge, Massachusetts and London, England. 3rd edition, 2009.
[29] P. Mankato. The Semantic Web - An Overview. [www] www.youtube.com, 2011. [Retrieved: September 2012].
[30] L. Watson. Swarming In Information Warfare. Edith Cowan University, Perth, Western Australia, 2002. [Retrieved: October 2012].
[31] H. Haken. Synergetics. Berlin: Springer-Verlag, 1983. (Cited in E. Bonabeau, M. Dorigo, and G. Theraulaz. “Swarm Intelligence From Natural to Artificial Systems.” New York: Oxford University Press, 1999).
[32] G. Nicolis and I. Prigogine. “Self-Organiization in Non-Equilibrium Systems.” NewYork, NY: Wiley & Sons, 1977. (Cited in E. Bonabeau, M. Dorigo, and G. Theraulaz. “Swarm Intelligence From Natural to Artificial Systems.” New York: Oxford University Press, 1999).
[33] J. Deneubourg, S. Goss, N. Franks, and J M. Pasteels. “The blind leading the blind: Modeling chemically mediated army ant raid patterns.” Journal of Insect Behavior, 2(5): pp. 719–725, 1989.
[34] B. Palace. Data Mining. Technology Note prepared for Management 274A, Anderson Graduate School of Management at UCLA [www] http://www.anderson.ucla.edu, 1996. [Retrieved: May 2012].
[35] S. Stefanov and V. Huang. “A semantic web based system for context metadata management.” Metadata and Semantic Research, 46: pp. 118–129, 2009.
[36] A. Vlachidis, C. Binding, D. Tudhope, and K. May. “Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature.” 2012. [www] hypermedia.research.glam.ac.uk [Retrieved: October 2012].

 ISBN: 978-1-61208-293-6