Resolving Terminology Heterogeneity in Digital Forensics Using the Web

[1,2]Nickson M. Karie[*], [1]H.S. Venter[†]
[1]Department of Computer Science, University of Pretoria,
Private Bag X20, Hatfield 0028, Pretoria, South Africa
[2]Department of Computer Science, Kabarak University,
Private Bag - 20157, Kabarak, Kenya
menza06@hotmail.com[*]
hventer@cs.up.ac.za[†]

Abstract: Frequency generators are devices that are widely used in the field of medicine in an attempt to effect chemical changes in the human body for the purpose of curing certain diseases. In addition, it is believed by many medical practitioners and researchers that frequencies can be generated, using sympathetic resonance to stimulate organ function or even to physically vibrate offending bacteria, viruses and parasites, resulting in their elimination from the human body. Digital forensics, which is rather a relatively new branch of forensic science, has attracted a wider array of people, such as computer professionals, law enforcement agencies and practitioners, that always need to cooperate in this profession. However, this has inflicted new problems with terminology heterogeneity within the domain, analogous to the offending bacteria, viruses and parasites in the human body, which needs to be resolved and/or eliminated.

This paper therefore, proposes a new method that uses the Web in an attempt to resolve terminology heterogeneity in the digital forensic domain. The proposed method is based on ‘terminology frequency’ which is automatically generated by the Web each time new information is added. We refer to this frequency as the Web Terminology Frequency (WTF) in this paper. Web search engines are however employed as tools that have the ability to capture the terminology frequencies. Finally, we show how the computed WTF is used to deduce a method coined as the ‘Terminology Heterogeneity Resolver’ (TE-HERE) in this paper.

Experiments conducted using the proposed TE-HERE method focuses on the digital forensic domain terminologies. However, in the authors’ opinion, the TE-HERE which is technically simple and does not require any human-annotated knowledge can as well be applied in other domains. This is because the findings presented in this paper indicate that the TE-HERE method produce influential results. TE-HERE is a novel approach to resolving terminology heterogeneity in digital forensics and constitutes the main contribution of this paper

Keywords: digital forensics, digital forensic terminologies, terminology heterogeneity, Web terminology frequency, Terminology Heterogeneity Resolver (TE-HERE), Web search engines

1. Introduction

The Web which dates back to 1989 (Gribble 2012, CERN 2012, WWW Foundation 2012) has grown to be such a vast entity where an astronomical amount of information is amassed. In addition, it is the largest semantic electronic database in the world. This "database" is available to all and can be queried using any Web search engine that can return aggregate hit count estimates for a large range of search queries (Cilibrasi and Vitànyi, 2007). New information is also added to the Web on a daily basis. To tap into this rich bank of information, Web search engines are the most frequently used tools to query for information related to a particular term (Karie and Venter, 2012). To the authors’ knowledge, there is so far no better or easier way to search for information on the World Wide Web than simply using Web search engines like Google. However, we do not dispute the existence of other techniques that can be used to search for and extract information from the Web. Therefore, in this paper we use the Web as a live and active electronic text corpus in an attempt to resolve terminology heterogeneity in digital forensics.

Heterogeneity and specifically, semantic heterogeneity is a problem that is not well understood in many domains. In addition, there is not even an agreement regarding a clear definition of this problem (Sheth and Larson, 1990). However, according to (Merriam-Webster Dictionary, 2012), heterogeneity refers to the state of being heterogeneous. Anything that is heterogeneous lacks uniformity. Heterogeneity can occur in a domain, for example, when the people involved have differences in perceptions. The use of different terminologies to describe exactly the same thing/object in digital forensics for example, causes terminology heterogeneity.

As a matter of fact, in the case where different digital forensics experts have to testify in a court of law, if evidence presentation is uniformly and correctly done using uniform terminologies, it is much more useful in apprehending criminals, and stands a much greater chance of being admissible in the event of a prosecution. Lack of uniformity in the usage of terminologies to communicate exactly the same evidence that convicts criminals might create loopholes for the attackers to evade responsibility. Therefore, resolving terminology heterogeneity can help in the above mentioned problems. This can further help create uniformity in communication, understanding, presentation and interpretation of domain knowledge and information.

As for the remaining part of this paper, section 2 discusses related work. In section 3 some technical background is explained, followed by a discussion of the proposed TE-HERE method in section 4. Experimental results are considered in Section 5, while conclusions are drawn in section 6 and mention is made of future research work

2. Related Work

Heterogeneity problems have been widely addressed at different levels by different researchers (Prasenjit and Gio, 2002). Various methods and models for resolving terminology heterogeneity have also been proposed. However, much of these methods and models are found in the domain of information sharing and particularly in interoperating databases (Xu and Lee, 2002). Existing research have also managed to identify different types of heterogeneity some of which include: structural semantic heterogeneity discussed by (Colomb, 1997) and schematic heterogeneity discussed by (Bishr, 1998). The biggest problem as discussed by (Colomb, 1997) lies in what can be called the fundamental conceptual heterogeneity. Fundamental conceptual heterogeneity occur when the terms used in two different ontologies have meanings that are similar, yet not quite the same (Sheth and Larson, 1990).

The concept of resolve terminology heterogeneity in ontologies is discussed by (Prasenjit and Gio, 2002). They propose a method that looks up in a dictionary or semantic network like WordNet (Miller, 1995) to determine similarities of words based on word similarity compound from a domain-specific corpus of documents. Their method however, focuses more on the use of information theory and hierarchical taxonomy such as the WordNet to resolve terminology heterogeneity in ontologies while in this paper we use the Web as a live and active text corpus (Xu et al, 2011) to resolve terminology heterogeneity in digital forensics.

In another paper (Sansonnet and Valencia, 2005) propose a method to solve semantic heterogeneity between software agents in an open world with a formalism for knowledge representation based on simplicial complexes. They propose an algorithm for solving terminological heterogeneity between knowledge bases formalized with the generalized simplicial notations. They further introduce the notion of an approximated mapping between heterogeneous terms. However, their work focuses more on software agents and knowledge bases while this paper focuses on digital forensic terminologies.

In their paper (Larab and Benharkat, 1996) introduces a schema integration method for federated databases based on a terminological reasoning approach. There method deals with the integration of terminologies that translate the export schemas (parts of DB^2 schemas which participate to the federation). However, the main focus of their method is in schema integration for federated systems based on a terminological reasoning approach while the current paper focuses on resolving terminology heterogeneity in digital forensics based on WTF.

In a paper by Serafini and Tamilin (2007) they address the problem of reasoning with instances of heterogeneously formalized ontologies. They build their approach upon the capability of mappings to enforce a propagation of concept membership assertions between ontologies. Their approach is grounded on a distributed description logic framework, which encodes ontologies as description logic knowledge bases and mappings as bridge rules and individual correspondences. They focus more on reasoning with instances of heterogeneously formalized ontologies while the current paper focuses on terminological heterogeneity in digital forensics.

In another paper Hakimpour and Geppert (2001) presents an approach to integrate schemas from different communities, where each such community is using its own ontology. Their approach focuses on merging ontologies based on similarity relations among concepts of different ontologies. They further present formal definitions of similarity relations based on intensional definitions. However, their approach concentrates on schema integration from different communities while we focus on terminology heterogeneity in digital forensics.

Another effort by Muresan et al, (2003) presents a two-step approach which they use to build a terminological database. Their work addresses obstacles to understanding results across heterogeneous databases brought about by the lack of ability to determine conceptual connections between differing terminologies. However, the problem with this approach is that it demands the construction of a terminological database while in our approach we use the Web which is considered the largest semantic electronic database in the world (Cilibrasi and Vitànyi, 2007).

There also exist other related works on resolving terminology heterogeneity; however, neither those nor the cited references in this paper have attempted to use WTF to resolve terminology heterogeneity in the way that is introduced in this paper. Our approach uses the Web and the Web search engines to resolve terminology heterogeneity in digital forensics. However we acknowledge the fact that the previous work on terminology heterogeneity has offered useful insights toward the development of the TE-HERE method in this paper. In the section that follows we present a detailed explanation of the technical background to aid in the understanding of the TE-HERE method discussed later in section 4.

3. Technical Background

Much of the theory explained in this paper is based on the WTF. Thinking of WTF for the first time can be very interesting. However, the theory of frequency has been used in many different domains for various purposes. Frequency generators for example, are widely used in the field of medicine in an attempt to effect chemical changes in the human body for the purpose of curing certain diseases (Frequencyrising, 2012).

Moreover, according to Kilgarriff (1997) frequency lists exist and are very useful representations of meaning for information retrieval, text categorisation, and numerous other purposes. In addition, frequency lists act as a representation of the full text which is susceptible to automatic objective manipulation. However, the full text which is very rich in information cannot be readily used to make for example, similarity judgments (Kilgarriff, 1997). Therefore, WTF is introduced in this paper in an attempt to resolve terminology heterogeneity in digital forensics.

According to (Cilibrasi and Vitànyi, 2007) Google events capture all background knowledge about the search terms concerned available on the Web. The Google event *x*, consists of a set of all Web pages containing one or more occurrences of the search term *x*. Thus, it embodies, in every possible sense, all direct context in which *x* occurs on the Web. This constitutes the Google semantics of the term *x*. For this reason, in all our experiments, the Google search engine was used.

Hit counts reported by Web search engines for a specified terminology *x* are useful information sources for this study and, as such, are used as input for computing the WTF. The hit counts of a search term query *x* is defined as the estimated number of Web pages containing the queried term *x* as reported by a Web search engine (Bollegala et al, 2011). However, the hit count may not necessarily be equal to the exact terminology frequency, because the queried term *x* may appear many times on a single Web page. However, for the purpose of this study, assuming that each Web page returned by the Web search engine contains a single instance of the searched term *x,* then it becomes clear that the total number of web pages (hit counts) reported by a search engine, can be equated to the estimated terminology frequency of the term *x* on the Web.

Note that, frequency is a general term used in different fields to define the number of occurrences of a repeating event *x* per unit time *t* (Bakshi et al, 2008). Calculating the frequency of any repeating event *x* can be accomplished by counting the number of times that event *x* occurs within a specific time period *t*, then dividing the count by the length of the time period as shown in equation 1.

$$Frequency\ (f) = \left(\frac{\text{Number of times an event } x \text{ occurs}}{\text{Length of the time period}(t)}\right) \quad (1)$$

Using equation 1 to compute the WTF, we replace the number of times an event x occurs with the hit counts reported by a Web search engine for the specified search term *x.* On the other hand, the length of the time period *t* is replaced by the time (in milliseconds) taken by the Web search engine to successfully execute the search query for the specified search term *x*. Substituting these values in equation 1 gives equation 2.

$$Frequency\ (f) = \left(\frac{\text{Hit counts for the specified search term } x}{\text{Length of the time period } (t) \text{in milliseconds}}\right) \quad (2)$$

The result obtained from equation 2 therefore gives a generalized WTF of the search term *x* with respect to the time (in milliseconds) taken by the Web search engine to successfully execute the search query for the specified search term *x.* Re-writing equation 2 in the context of WTF gives equation 3.

$$WTF = \left(\frac{\text{Hit counts for the specified search term } x}{\text{Length of the time period } (t) \text{in milliseconds}}\right) \qquad (3)$$

However, for the purpose of this study, the following notations are adopted.

f(x) denotes the estimated hit counts returned by a Web search engine for any specified search term *x* on the Web.

f(t) denotes the time period *t* in milliseconds taken by the Web search engine to successfully execute a search query for the given search term *x*.

Replacing these notations in equation 3, gives equation 4.

$$WTF = \left(\frac{f(x)}{f(t)}\right) \qquad (4)$$

Equation 4 therefore is used in this study as the basis for computing the WTF for any given digital forensics terminology *x* on the Web. The proposed TE-HERE method also utilizes the computed WTF in resolving terminology heterogeneity and is explained in the section that follows.

4. The Proposed TE-HERE Method

Cooperation among different people working in a domain is inevitable and often involves handling of information from diverse sources, analysing it and further presenting it to other stakeholders. Information sources however, cannot be predetermined because they may be autonomously created and maintained (Prasenjit and Gio, 2002). The owners of the information sources at times may as well prefer to maintain their autonomy.

However, effective cooperation among different people in any domain presupposes that information from different sources should be harmonised in such a way as to create uniformity and common understanding in the domain. The harmonisation process however, can be very costly and close to impossible if done manually; especially when the people involved have differences in background and perceptions on the meanings and usage of certain domain terminologies.

Therefore, we propose in this paper a method coined as the **T**erminology **He**terogeneity **R**esolver (TE-HERE) which utilizes WTFs to resolve terminological heterogeneity.

With reference to equation 4, many of the Web search engines deliver search results within fractions of a second (milliseconds). In addition, the search query execution time even for the same search term *x* using different search engines might in many cases be different. Some search engine deliver search results faster than others. Therefore, to eliminate on these variations in the search query execution time, the authors introduce an assumed search query execution time of one second (1000 milliseconds) across all search engines. Thus, for computing the WTF, the search query execution time used is one second irrespective of the search engine used. This is done by multiplying the result of equation 4 by 1000 milliseconds equivalent to 1 second.

As a concrete example, let the specified search term *x* be ‘Digital Evidence’ using the Google search engine as of 14th June 2012, the search query executed returned 103,000,000 hits in 0.16 seconds (160 milliseconds). Using equation 4 to compute the WTF gives a value of **643,750**. However, assuming an execution time of one second, the results of equation 4 is further multiplied by 1000 as shown in equation 5.

$$WTF = \left(\frac{f(x)}{f(t)}\right) * 1000 \qquad (5)$$

From equation 5, the WTF changes to **643,750,000**. For this reason infer from this calculation that irrespective of the search engine used for the specified search term *x,* if all other factors remain constant, then the computed WTF would be **643,750,000**. Equation 5 is therefore used in this study to define the proposed TE-HERE method and can be re-written as equation 6.

$$\text{TE-HERE} = \left(\frac{f(x)}{f(t)}\right) * 1000 \qquad (6)$$

Equation 6 therefore, defines the TE-HERE, a new method for resolving terminology heterogeneity in digital forensics using the Web and Web search engines. The experimental results obtained using the proposed TE-HERE method was found to be influential and are discussed in the section that follows.

5. Experimental Results

As mentioned earlier, the theory of frequencies has been used in different domains for different purposes. However, in this study, WTF is introduced and used to resolve terminology heterogeneity in digital forensics. Terminologies which produce the highest TE-HERE values are deemed to have the highest frequency. Therefore, it should be possible to use the realized frequency to influence domain members to adopt the use of certain terminologies during communication and presentations of domain knowledge and information. This also implies that, terminologies with the highest TE-HERE values can be adopted for use, for example, in a court of law or civil proceedings where uniformity in the understanding and interpretation of evidence information by all stakeholders is a priority.

While the theory discussed in this paper is rather intricate, the resulting method is simple enough. Knowing that there exists terminology heterogeneity in digital forensics, the computed TE-HERE values of the terms in question as defined by equation 6 can be used as a quick guide to resolve the terminology heterogeneity.

Given any two digital forensic terms, for example, x_1 and x_2 used to refer to the same thing; we find the number of hit counts for the terms, denoted as $f(x_1)$ and $f(x_2)$. We also note the time period in milliseconds $f(t_1)$ and $f(t_2)$ taken by the Web search engine to successfully execute the search term queries.

As a concrete example, let the search term x_1 be 'Digital evidence' and search the term x_2 be 'Electronic evidence'. Using the Google search engine, the hit counts reported for the term x_1 and x_2 as on 14 June 2012, it follows that:

"Digital evidence" $f(x_1)$ =103000000 and $f(\mathrm{t}_1)$ = 160 milliseconds (0.16 sec)
"Electronic evidence" $f(\mathrm{x}_2)$ =34900000 and $f(\mathrm{t}_2)$ = 330 milliseconds (0.33 sec)

Substituting these values in equation 6 gives the following TE-HERE values 'Digital Evidence' = 643,750,000 and 'Electronic evidence' = 105,757,575.8. Since Digital Evidence has the highest TE-HERE value, it simply means that it has the highest frequency of usage and thus preferred by many as compared to Electronic evidence. It can also mean that, in case of a digital forensics investigation, the term 'Digital Evidence' can be used in the place of 'Electronic Evidence' without misleading the receivers of such information.

To further analyse the performance of the proposed TE-HERE method, we conducted two sets of experiments. First we use a data set proposed by Prasenjit and Gio (2002). Secondly, the proposed TE-HERE method is tested using digital forensics domain terms to measure its performance against selected terms. These two experiments are discussed in the two sub-sections that follow respectively.

5.1 The Mitra and Wiederhold Data Set

To test the performance of the proposed TE-HERE method, we use a data set proposed by Prasenjit and Gio (2002) in Table 1. This data set was used in two different ontologies. However, an analysis of the different terms showed that they were used to refer to the same thing. The input to the TE-HERE method is therefore the reported Google hit counts $f(x)$ and the time period $f(t)$ for any of the terms in question. The TE-HERE method works by computing the TE-HERE value for each of the terms (see Table 1) using equation 6. Given any of the terms pairs x_1 and x_2, the associated computed TE-HERE values determine their frequency as seen from table 1.

Table 1: Mitra and Wiederhold data set, *Original source:* (Prasenjit and Gio 2002*).*

Term (x_1)	**TE-HERE Values**	**Term (x_2)**	**TE-HERE Values**
Passenger	885185185.2	Passenger	885185185.2
Cargo	**1920833333**	Payload	**92727272.73**
Departure Time	24565217.39	Time	38066666667
Arrival Time	38391304.35	Time	38066666667
Arrival City	9148148.148	Destination	2495652174
Name	27242424242	Location Name	25250000
Departure City	17842105.26	Origin	3052000000
Airport	2637931034	Airforce Base	2182352.941
Flight	2496296296	Sortie	665217391.3

The TE-HERE Values shown in Table 2 depicts the estimated frequency of the digital forensic terms in question. Table 1 on the other hand, was used mainly for the purpose of testing the proposed TE-HERE method with terminologies not necessarily originating from digital forensics. This was done to provide a clear picture of the performance and accuracy of the TE-HERE method.

From Table I, column 1 and 3 shows the terms used while column 2 and 4 indicate the TE-HERE Values computed using equation 6. For example, the terms ‘Cargo’ and ‘Payload’ (See Table 1) with TE-HERE Values of 1,920,833,333 and 92,727,272.73 respectively, depicts the performance of the TE-HERE method. In this case, because Cargo has the highest TE-HERE value, it can be adopted in place of Payload without distorting information. This is also depicted in the other columns of Table 1. From this experiment, it is clear that using the TE-HERE method; we can generate satisfactory terminology frequency results. However, it is also possible to get false positives; though in a significant number of cases the TE-HERE values were satisfactory. Therefore, we suggest that for any application, if the results are not satisfactory the domain expert can decide based on the computed TE-HERE Values which of the terms in question can be adopted for use. For example, in the case of ‘Departure Time’ and ‘Time’ (see Table 1) a decision can be made to decide on which of the terms can be adopted for use.

5.2 Digital Forensics Terminologies

In Table 2, a part of the experimental findings is presented using selected digital forensics domain terms. Each term enclosed in double quotes is used as a single Google search term with the reported hit counts denoted in Table 2 as *f*(*x*) and the associated execution time period denoted as *f*(*t*). The computed TE-HERE value shows the measures obtained to ascertain the performance of the TE-HERE method. The selected digital forensics terms used are: 'Digital Evidence', 'Electronic evidence', 'Multimedia evidence' and 'Digital and multimedia evidence’.

The authors found that these terms are mostly used in discussions that involve the digital forensics investigation and evidence presentations in either a court of law or civil proceedings, hence the motivation for the experiment indicated in Table 2. In all the experiments conducted, the TE-HERE method produced influential results as can be seen from Table 1 and 2 respectively.

To determine the TE-HERE value of the search terms in Table 2, the proposed TE-HERE method was used. The first column of the table show the sampled digital forensics terminologies and their corresponding hit counts values *f(x)* shown in column two, the time taken in milliseconds by the search engine to execute the search query *f(t)* is shown in column three and finally the TE-HERE values are presented in the last column.

Table 2: Experimental Findings of Digital Forensic Terminologies using the TE-HERE method

Digital Forensics Terminologies	**Hit Counts *f(x)***	**Time in milliseconds *f(t)***	**TE-HERE Values**
Digital and multimedia evidence	17300000	240	72083333.33
Multimedia evidence	61400000	210	292380952.4
Digital Evidence	103000000	160	643750000
Electronic evidence	34900000	330	105757575.8

In the case of a need to resolve terminology heterogeneity in the digital forensics domain and/or any other domain for example, the proposed TE-HERE method can be used. Since the term ‘digital evidence’ has the highest TE-HERE value of **643,750,000** it simply means that, ‘digital evidence’ is widely used and probably preferred by many stakeholders.

Therefore, this value taken as the global estimated terminology frequency can be used to influence other members to adopt the use of the term ' digital evidence ' in all communications and presentations involving digital forensic evidence either in a court of law or civil proceedings. This will as well create uniformity in communication, understanding and interpretation of domain knowledge and information.

The performance of the proposed TE-HERE method is further backed up by a graphical representation of the Table 2 results shown in Figure 1.

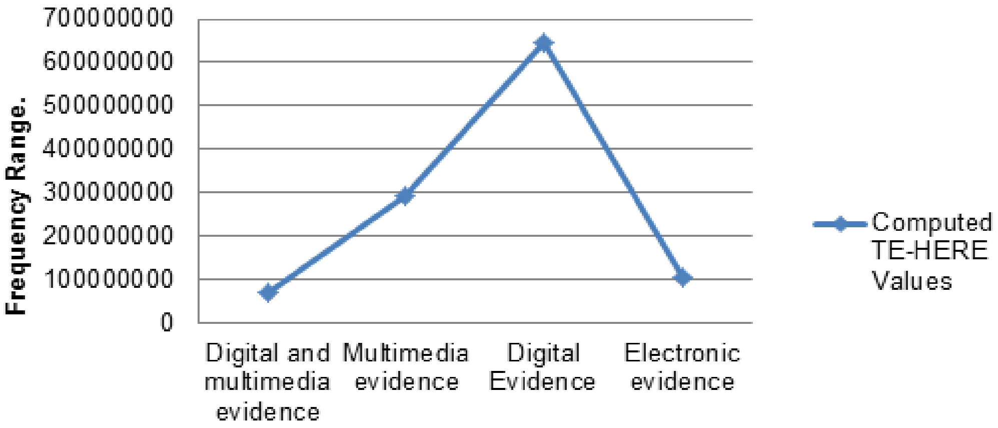

Figure 1: Graphical Representation of the TE-HERE Values from Table 1.

In this paper, we have adopted a relatively technical simple and new method for resolving terminology heterogeneity in digital forensics coined as the "Terminology Heterogeneity Resolver" (TE-HERE). This method uses the Web as a live and active text corpus, comprising automatic generated WTF. In addition, Web search engines are employed in this study as tools that have the ability to capture the terminology frequency.

To the best of the authors' knowledge, there exists no other experiments in digital forensics similar to the one explained in this paper, that can be used as a baseline to judge the performance of the proposed TE-HERE method thus found it hard to benchmark the measures produced by the TE-HERE method for the specified digital forensic terminologies. This is, therefore, a novel approach of using the Web and the Web search engines to resolve terminology heterogeneity in digital forensics.

The advantage of using the TE-HERE method is that, there is so much of the data involved (The entire Web) thus; TE-HERE values can be generated with respect to the thousands of data points available on the Web. The TE-HERE method is also economical and technically simple because it utilises the freely available information on Web and the Web search engines. In the case of digital forensics, for example, the TE-HERE method can be used to influence members to adopt the usage of certain domain terminologies.

6. Conclusion

The problem addressed in this paper was that of terminology heterogeneity in the digital forensics domain where different stakeholders use different terminologies to describe or refer to exactly the same thing or object, resulting in the lack of uniformity in communication, understanding and interpretation of domain knowledge and information. Knowing that digital forensics is relatively a new field, addressing terminology heterogeneity at an early stage can help resolve both the present and future terminology heterogeneity problems, thereby creating a lasting uniformity in the domain.

Finally, the proposed TE-HERE method was found to generate influential results making it a method to consider as a quick guide for resolving terminology heterogeneity. This is backed up by the fact that the TE-HERE method is technically simple and economical. Though the initial experiments done were based on the digital forensics domain terminologies, the authors believe that the TE-HERE method can be applied in other domains as well. As part of the future work, the authors are now planning to use the TE-HERE method to build an automated model for resolving terminology heterogeneity and even more as a way towards resolving semantic disparities in the digital forensics domain. However, more research needs to be done so as to improve on the performance and accuracy of the TE-HERE method.

References

Bakshi, U.A., Bakshi, A.V., and Bakshi, K.A. (2008), Time and Frequency Measurement, Technical Publications, (Electronic Measurement Systems.) First Edition:2008 pp. (4–1). [online], http://books.google.co.za/books?id=jvnI3Dar3b4C&pg=PT183&hl=en#v=onepage&q&f=false

Bishr Y.A., (1998), Overcoming the Semantics and Other Barriers to GIS Interoperability. International Journal of Geographic Information Science, Vol. 12, No. 4, pp 299-314.

Bollegala, D., Matsuo, Y. and Ishizuka, M. (2011), A Web Search Engine-Based Approach To Measure Semantic Similarity Between Words, IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 7, pp 977-990.

CERN, (2012) How the web began [online], http://public.web.cern.ch/public/en/about/webstory-en.html

Cilibrasi, R.L. and Vitànyi, P.M.B. (2007), The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, pp. 370–383.

Colomb, R.M. (1997) Impact of Semantic Heterogeneity on Federating Databases, The Computer Journal, Vol. 40, No. 5 pp 235-244.

Frequencyrising, (2012) Bio Frequency Generator, [online], http://www.frequencyrising.com/frequency-generator.html

Gribble, C. (2012), History of the Web Beginning at CERN, [online], Hitmill, http://www.hitmill.com/internet/web_history.html

Hakimpour, F. and Geppert, A. (2001), Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach, Proceedings of the of the International Conference on Formal Ontologies in Information Systems, pp 297—308.

Karie, N.M. and Venter, H.S. (2012), Measuring semantic similarity between digital forensics terminologies using web search engines, Proceedings of the Annual Information Security for South Africa (ISSA, 2012) Conference. pp.1-9,Sandton, Johannesburg. Published online by IEEE Xplore®

Kilgarriff, A. (1997), Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. Proceedings of the AISB Workshop, Falmer.

Larab, O. and Benharkat, A., (1996), Resolving Semantic Heterogeneity in Databases with a Terminological Model: Correspondence Refinement. Proceedings of the International Workshop on Description Logics, Cambridge, MA, USA. pp.150-154.

Merriam-Webster Dictionary, (2012) [online], http://www.merriam-webster.com/dictionary/heterogeneity

Miller, G.A. (1995), “WordNet: A Lexical Database for English”, [online], PrincetonUniversity, http://wordnet.princeton.edu/

Muresan, S., Popper, S.D., Davis, P.T., and Klavans, J.L. (2003), Building a Terminological Database from Heterogeneous Definitional Sources. Proceedings of the annual national conference on Digital government research 2003.

Prasenjit, M. and Gio, W. (2002), Resolving terminological heterogeneity in ontologies, Proceedings of the ECAI-02 Workshop on Ontologies and Semantic Interoperability. Lyon, France, July.

Sansonnet, P. and Valencia, E. (2005), Terminological Heterogeneity Between Agents Using a Generalized Simplicial Representation, EUMAS 2005- Proceedings of the Third European Workshop on Multi-Agent Systems, Brussels, Belgium, pp 363-374.

Serafini, L. and Tamilin, A., (2007), Reasoning with Instances of Heterogeneous Ontologies. [online], Data & Knowledge Management Group, http://ceur-ws.org/Vol-314/52.pdf

Sheth, A.P. and Larson, J. (1990), Federated database systems for managing distributed Heterogeneous and autonomous databases. ACM computer Surveys, Vol 22 No. 3, pp 183 – 236.

WWW Foundation, (2012), History of the Web, http://www.webfoundation.org/vision/history-of-the-web/

Xu, Z. and Lee, Y.C. (2002), Semantic heterogeneity of geodata, Proceedings of the Symposium on geospatial theory, processing and application, Ottawa.

Xu,Z., Luo, X., Yu, J. and Xu, W. (2011), Measuring semantic similarity between words by removing noise and redundancy in web snippets. Concurrency and Computation: Practice and Experience , Vol. 23 No. 18, pp 2496-2510.