Proceedings of the e-Skills for Knowledge Production and Innovation Conference 2014

Cite as: Agbaje, M. O., Awodele, O., & Ogbonna, A. C. (2014). Big data, audience measurement and digital watermarking: A review. *Proceedings of the e-Skills for Knowledge Production and Innovation Conference 2014, Cape Town, South Africa,* 17-28. Retrieved from http://proceedings.e-skillsconference.org/2014/e-skills017-028Agbaje840.pdf

Big Data, Audience Measurement and Digital Watermarking: A Review

Micheal Agbaje, Oludele Awodele, and A. Chibueze Ogbonna
Babcock University, illisan- Remo, Ogun State, Nigeria

agbajeolugbenga@gmail.com; delealways@yahoo.com; acogbonna06@yahoo.com

Abstract

The objective of this paper is to provide some background to those interested in big data, audience measurement and digital watermarking. These technologies are currently in trend and are linked together. Big data provides techniques for analysis of complex and large data produced by audience measurement: radio, TV, Internet, Newspaper, religious and Education while digital watermarking is useful in counting audience to produce accurate data. The paper uses exploratory technique to achieve its objectives. We discuss some introductory concepts on the three types of technologies and some alternative methods employed for each service and also proffer a way forward using digital watermarking for a solution. The paper concludes by looking at the usage of the technology in Africa present the way forward into the future.

Keywords: Big data, audience measurement, digital watermarking

Introduction

Big data, according to Wikipedia, is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications while *audience measurement*, as the term is commonly used, refers to regular assessments of the size and composition of media audiences (Muller, 2001).

The interaction of connected consumer electronics and digital media is creating vast and limitless amounts of user data, now commonly coined "Big Data." Many new business models are forming around these data, but for advertising, data have always been at the core of its business. Data of all varieties and volumes enable stakeholders to invest proportionally to the value of their media assets; whether TV commercials, video programs, gaming apps, online publications or advertising messages (Way, 2014).

Big Data create new ways that stakeholders can measure the audience. Traditional targeting has long been audience-based using demographic, geographic and time-based information. However, user data originating from content consumed on an Internet-enabled device is much more granular in nature -- more data points are collected, supporting highly refined audience-targeting methods, including both audi-

This paper accepted by executive review

ence-based and interest-based methods (Way, 2014). Audience-based data and targeting practices segment consumers based on who they are, the devices they use, and the media content they consume. Interest-based measures support techniques that pinpoint and engage consumers based on their interests and preference (Way, 2014).

Digital watermarking is a technology that embeds information, in machine-readable form, within the content of a digital media file (for example, a movie, song, or photograph). Digital watermarking is the process of embedding information into digital material in such a way that it is imperceptible to a human observer but easily detected by computer algorithm (Mooney & Keating, 2005; Seitz, 2005; Serra-Ruiz, & Fallahpour, 2010).

The accuracy of the data is very important to a good decision making. There is various method of collection of data. Digital watermarking can serve as a means of accurate data collection mechanism thus the link between the three technologies giving birth to the title of the paper.

We perceived a state of neglect to this aspect in most African countries with the exception of South Africa which has developed an audience measuring culture. In Nigeria and most of the African countries there has not been a deliberate structure on ground to measure audience except for individual companies relying on record of sales and some other form of marketing research. The closest to it is by monitoring the number of callers that called on a program but no dedicated peoplemeter or panel for any TV or radio. Also the popular webometric data for measuring how many people access institutional data is quite popular and is being used to rate universities here in Nigeria. But what other usage and analysis of the data generated from this measurement over a long period is quite unknown.

Therefore the objective of the paper is to bring to fore the relationship between the emerging technology of audiences measurement, big data and digital watermarking the various uses. This paper used an exploratory technique in its approach by studying works of other authors and exposing the need to embrace the subject matter in view of their importance to the society.

The Arab News (2014) reported the training of Lectures on big data and Analytics. They said the global demand for big data jobs is currently in the millions worldwide. This is between 1- 1.5 millions in the US alone. Institutions can set a rebuilt of their curriculum to the current trends. Part of the Saudi e-government vision is to digitize and automate all government transactions and correspondence across all ministries. This is predicted to generate more and more data which creates room for the adoption of big data technology.

Literature Review

Concepts of Big Data

What is big data?

Big data, according to Wikipedia, is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured (SAS Institute, 2014).Big data may become important to business and society such as the internet is today. This is because more data may lead to more accurate analysis.

Big data analytics

Big data analytics is the process of examining large amount of data of a variety (big data) to uncover hidden patterns, unknown correlation and other useful information. Big data technologies: NoSQL databases, Hadop, and MapReduce.

The definition of big data can be done using three Vs: volume, velocity, and variety

Volume: Many factors contributed to increase in volume:

- Transaction based data stored through the years.
- Unstructured data streaming in from social media.
- Increasing amount of sensors and machine-to- machine data being collected.

Velocity: Data is streaming at unprecedented speed and must be dealt with in timely manner. RFID tags and smart metering are driving the need to deal with data in near real-time.

Variety: Today's data come in all formats: structured, numeric data in traditional databases, unstructured text documents, email, video, audio stock ticker data, and financial transactions (SAS, 2014).

Other dimensions of big data are:

Variability: Data flow can be highly inconsistence with periodic peaks

Complexity: Today's data comes from multiple soured. It is important to link sources, match cleanse and transform data across systems. It is also necessary to connect and correlate relationships hierarchies and multiple data linkages or data can spiral out of control.

Benefits of big data

The issue is not collecting large amount of data. It is what you do with it that matters. Organisation should be able to take data and analyse to find answers that produce cost reduction, time reductions, new product development, and smart business decision making. Combining big data and high powered analytics, it is possible to:

- Determine root cause of failure, issues defects in real-time, potentially save billions of dollars annually.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Quickly identify customer who matter most.
- Use clickstream analyses and data mining to detect fraudulent behaviour.
- Generate retail coupons at the point of sale based on the customers' current and past purchases.
- Optimise routes for thousands of delivery packages while they are on the road, e.g., UPS.

Big data in action

United Parcel Services (UPS)

UPS stated capturing data and track variety of transaction in the 1980.They now track data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking request by customers. The company stores 16 petabyte of data.

Sloan Digital Sky survey (SDSS)

The Sloan Digital Sky Survey (www.sdss.com) is the most ambitious astronomical survey ever taken. It started collecting data in 2000.It has collected in a few weeks' data more than all data collected in history of astronomy. At the rate of 200GB per night they have amassed 140 terabytes of information ("Sloan Digital Sky Survey," n.d.).

DNA

Decoding of human genome originally took 10 years to process, now it can be achieved in less than a week.

NASA

The centre for climate simulation (NCCS) stores 32 petabytes of climate observation on the Discover supercomputing cluster.

US Government

In 2012, the Obama administration announces Big Data Research and Development Initiative, which explore how big data could be used to address important problem faced by government. Big data analysis played a role in Obama's successful 2012 re-election campaign. US government owns six of the ten the most of the powerful supercomputers in the world (Hoover, 2012; "How bid data…," 2013; Kalil, 2012)

Big data sofware

- **Hadoop-** Apache Foundation
- **MongoDB**- MongoDB Inc
- **Splunk**- Splunk Inc. etc.

Concepts of Audience Measurement

Audience measurement refers to regular assessments of the size and composition of media audiences (Muller, 2001). The concept of audience measurement stated as far back as the 1930 in America when readership of newspaper matters to the distribution and till the present internet era. This was pioneered by Nielsen Inc. In the era of digital broadcasting and with many new ways of watching TV content, accurate audience measurement has become more difficult. Audience Measurement services must now report more accurately and reliably from a larger number of channels, delivered through a fast-changing and diverse mix of broadcast platforms, and consumed either in real time or time-shifted mode.

Audience-based data and targeting practices segment consumers based on who they are, the devices they use, and the media content they consume. Interest-based measures support techniques that pinpoint and engage consumers based on their interests and preference (Way, 2014).

Uses of audience measurement

The major impetus for audience measurement is advertising. By the late nineteenth century most newspapers and mass circulation magazines had begun to sell space to advertisers. The value of space was determined largely by the number of readers who would see each publication, and by extension each ad. Paid circulation served as a reasonable surrogate for the size of the readership (Muller, 2001).

Today, advertising is a multi-billion dollar business. Audiences are bought and sold like commodities. Advertising expenditures are typically guided by audience measurement and the cost of

reaching various audience segments. In a business world increasingly interested in target marketing, research firms have been called upon to produce ever finer demographic distinctions, as well as data on lifestyles and product purchases.

The fact that advertising is the major source of revenue for several forms of media (including broadcasting, newspapers, and magazines) has embedded audience measurement in the operation of these industries. Obviously, the system places a premium on audiences that will be attractive to advertisers, either by virtue of their sheer size or desirable composition. Content, therefore, is evaluated with an eye toward its audience-making potential (Muller, 2001).

Sources of error in audience measurement

There are four sources of error in audience measurement: sampling, non response, response, and processing. The first three are, again, problems commonly associated with survey research. The last includes a variety of issues that have to do with bringing a saleable research product to market (Muller, 2001).

Concepts of digital watermarking

Digital watermarking is the process of embedding information into digital material in such a way that it is imperceptible to a human observer but easily detected by computer algorithm (Megías, Serra-Ruiz, & Fallahpour, 2010; Seitz, 2005). A digital watermark is a transparent, invisible information pattern that is inserted into a suitable component of the data source by using a specific computer algorithm (Katzenbeisser & Petitcolas, 2000; Petitcolas, Anderson, & Kuhn, 1999). Digital watermarks are signals added to digital data (audio, video, or still images) that can be detected or extracted later to make an assertion about the data.

Benefits of digital watermarking for audience measurement

- Accuracy and detailed detection logs allows the reporting of the content being watch, channel, airing time and distribution network.
- Can be used for radio, television, internet video and podcast audience measurement applications.
- ID tags or payloads enable transmission of significant amounts of data, providing superior audience granularity.
- Can be integrated into legacy audiometer to minimize change for panellists and audience operators.
- Allow improved live broadcast reporting as well as excellent time-shifted measurement.

Methodology

Audience-Based Measurement and Platforms

A number of factors such as transmitter power, local geography, station programming, wavelengths, and numerous other factors are known to influence the size of the audience (Bornman, 2008). Also, media consumers are not glued to one device, so an immediate concern is one of audience duplication, i.e., spending the time and money only to target the same user multiple times.

Despite industry demand, there has been difficulty developing a standard that accounts for the overlap of traditional media and new media. A true cross-platform measurement approach quantifies unduplicated reach across broadcast and cable TV, the Internet, and mobile apps and Web properties. However, perpetual device fragmentation, differing data collection methods and re-

porting metrics, and the lack of an industry-backed system to standardize the process stifle industry efforts (Way, 2014).

- **Circulation Figures:** Collected by newspaper and magazine producers, based on copies sold or through marketing agencies/providers. Supplies the numbers of newspapers/magazines sold for a given period within a given geographical area.
- **Focus Groups:** Predominantly film, and sometimes magazines, uses focus groups. A feature film will be produced. If there is some doubt as to the impact or audience for the film, it will be shown to selected consumers and they will be asked to complete a questionnaire. Based on the results of the questionnaire, the film may be modified, for example, a new ending created or further editing to speed action
- **Ratings: Nielsen Media Research Television Ratings** represent the industry currency for television audience measurement in most developed countries. Rating numbers are the average audience rating or the percent tuned to a particular programme during the average minute. A TV rating only measures how many people had the opportunity to watch. Therefore, programmes that have the larger audience are, by definition, the most successful ones.
- **PeopleMeters:** PeopleMeters are TV top-top boxes with remote controls used by all members in the selected household, each with their own codes. The data is collected every fifteen minutes and sent daily, every night, via a telephone line, to Nielsen. Overnight ratings data is available every day. See Figure 1:

Figure 1: **Handset and display unit of the peoplemeter used in South Africa (**Source: AGB Nielsen Media Research).

Nielsen has consequently embarked on one of the world's most comprehensive approaches to channel detection. Where a broadcaster agrees to cooperate, Nielson will take the initiative to place an invisible or inaudible signal in the channel's video or audio stream to permit the measurement of the broadcasts. Even when no active code is embedded, digital broadcasts can be iden-

tified by taking the video or audio signatures collected by meters and matching them to a reference data base of all possible signatures. This combined methodology patented worldwide by Nielsen make the identification of channel-specific viewing possible within an analogue, digital or mixed analogue-digital environment possible even without the cooperation of operators (Bornman, 2008).

Audience of Print Media: Newspapers and Magazines

Similar to most other media industries, editors and organisations involved in the publication of newspapers and magazines operate in two markets. The first is the market for the selling of copies. With regard to this market, readership data provide editors and circulation departments with information on the relative "success" of the publication in attracting the size and profile of the audience aimed for.

The audience size of a newspaper or magazine is usually measured in terms of the average issue readership, that is, the number of different people that reads a particular issue averaged across issues (coverage). Here it is important to point out that each copy of a particular issue could potentially have several readers. It is, however, insufficient to categorise people either as readers or non-readers of a particular newspaper or magazine. It is also necessary to establish the regularity or frequency of their reading (frequency). Frequency is usually indicated by the probability of contact with a particular issue. In the absence of electronic metering devices for measuring readership, readership research is mostly dependent on more traditional research methodologies and techniques.

Radio Audience

The unique nature of the radio as a broadcast medium presents problems to audience research which are in many ways not only different to researching television audiences, but also make it relatively more complex and difficult It is, in fact, the advantages of radio as a medium − the fact that the medium is mobile and allows people to go on with their daily activities instead of requiring everything to come to a standstill − that make it difficult to measure radio listening:

The following techniques are employed in measuring radio audiences.

- Surveys in which respondents are questioned with regard to what they usually listen to, when they usually listen and how often they listen to particular programmes, namely their radio listening habits.
- Diaries.
- Metering – Audio metering was introduced before its applications for television. However, the growth in radio mobility (e.g., the development of car radios, the explosion in the availability of small portables) and the rise of multi-set ownership led to the demise of these systems of radio metering.

TV Audience Measurement

Audience measurement in television viewership is intended to collect information on the audiences watching a specific television program at a particular time. To accurately measure TV audience, a panel of representative audiences must be selected judiciously so that it accurately represents the entire target audience group. However, it is hard to secure a proper number of target audiences due to the expensive and cumbersome installations of measurement equipments.

Lim et al. (2013) resolved this issue in panel selection by proposing a novel television audience measurement framework (Figure 2) using pervasive smart devices such as a smartphone. In the proposed framework, a short audio signal from a television set is recorded by a personal smart

device and is sent to an audio matching server for the identification of the television program shown by the television set. For effective identification, they propose an accurate audio matching algorithm based on spectral coherence and efficient implementation techniques that exploit the inherent parallelism in the algorithm (Lim, Choi, Nam, & Chang, 2013).

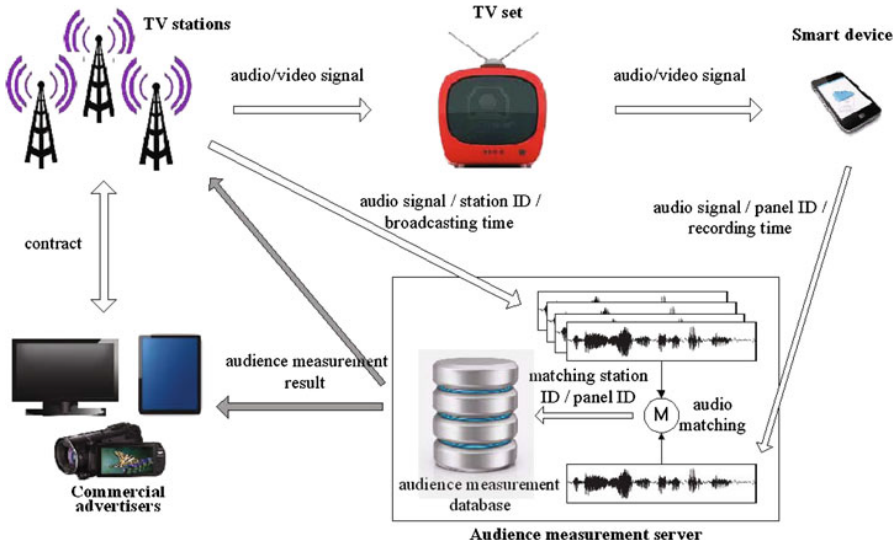

Figure 2: Proposed audience measurement (TV) framework by Lim et al, 2013.

Internet Audience Measurement

The arrival of Internet was around 1990 and by 1994 consumers was not so familiar with the internet. Consumer's closest experience was with the commercial online services, CompuServe, Prodigy, and the rapidly growing America Online. The Internet addressable email system was only just becoming available on some of these systems in the later part of 1994 (Coffey, 2001). The Prodigy service was the first of the "big three" online services to provide a workable Internet interface for its subscribers.

External audience measurement of the commercial online services was not in demand at the time. The online services had an excellent subscriber counting mechanism in the form of subscription accounts with which to keep track of their performance. Standard accounting and internal metrics met most of their measurement needs. A simple periodic survey was sufficient to gauge relative market share, and none but the curious were otherwise interested in the size and composition of each of the firm's audiences, principally because no third-party significant decisions were being made based on the estimates (Coffey, 2001).

By 1995, however, Prodigy had launched its web interface as had CompuServe and America Online. PC ownership and commercial online service subscriptions were rising rapidly, very well in large part to AOL's windows-based easy to use interface Technical advances were providing more and more people with easier access to the Internet. Investment began flowing into Internet companies, particularly in Silicon Valley, which fueled further growth and media coverage (Coffey, 2001).

Uses of Internet audience measurement

Internet audience measurement is used for three main purposes:

- Self-promotion. It is important for organizations to be able to make claims about the size and growth of their audiences or technologies. While internal records are very valuable,

they are often not audited for external use and it is difficult to compare the results with those of competitors (Coffey, 2001).

- To support advertising planning, buying, selling and posting. Organizations offering Internet media opportunities to advertisers or their agencies, use audience measurement data to help position and sell the inventory. This is the same role that television ratings, radio ratings, and magazine audience estimates play for their respective media. The planner uses ratings data to sort through the many different options available, to identify those that are better values for the target audience (Coffey, 2001).
- Strategic planning. The data are a treasure trove of information once properly mined. Knowing the patterns of consumer behavior, how consumers interact with a particular site or group of sites, can help site managers make decisions that improve the traffic flow and objective of the site tremendously.

Methods of internet audience measurement

There are three primary methods of Internet audience measurement in use today, with each having several variations. These are:

- Measurement from a sample of users who are metered (electronic measurement)
- Measurement from a sample of users who are surveyed (recall measurement)
- Measurement from analysis of server log files or their equivalents

Survey from a sampled user: These studies draw a sample of Internet users and then query the respondents through standard survey methods. This could be done through telephone, in-person, mailed, or web-based interviews. The advantages of this approach is that pertinent, definitive detail about individual users can be captured, such as age, sex, income, geography, and so on. The survey method for measuring a specific site's audience is frustrated by three factors.

- Over- claiming is a significant problem for very well known properties, as sites with very high brand awareness are often claimed when no actual usage took place.
- Social desirability (or undesirability) can have a powerful influence on claimed usage. Visitation to adult content sites, for example, will naturally be under-reported, especially when a live interviewer is involved.
- Since usage estimates will be based on recall, naturally occurring errors in memory will affect the results (Coffey, 2001).

Proposed Audience Measurement using Digital Watermarking

The Digital Watermarking Alliance describes the issue of audience measurement as follows (“Audience measurement, n.d.)

> In this new media world of insatiable content consumption, audience measurement is becoming more and more critical. Beyond the hard numbers of how many people are accessing a program, understanding who is watching, how they engage with the content, when, where and through which media is essential for content providers, advertisers and broadcasters to better tailor their offerings and maximize impact. The proliferation of devices and networks for watching content, the multitude of ways to watch such content and the changing habits in viewing content, such as PVR and catch-up TV services, is making audience measurement far more complex than ever before. Audience Measure-

ment services must now report more accurately and reliably from a larger number of channels, delivered through a fast-changing and diverse mix of broadcast platforms, and consumed either in real time or time-shifted mode.

Digital watermarking is proposed as a solution to accurate data gathering. Digital watermarking embeds a unique identifier into media content while being distributed or prior to distribution, making content and corresponding broadcasters instantly identifiable. Using specialized software able to retrieve, analyze and report the data, digital watermarking allows the precise identification of content and broadcasters.

The technology works by inserting digital data, imperceptible to the human ear, into each program's audio track. The digital ID contains information about the channel that broadcast the program, the airing time and, if relevant, a content identifier. Audiometers installed in panellists' homes read the data, collect the information and send them daily to a central database for processing and accurate reporting as shown in Figure 3

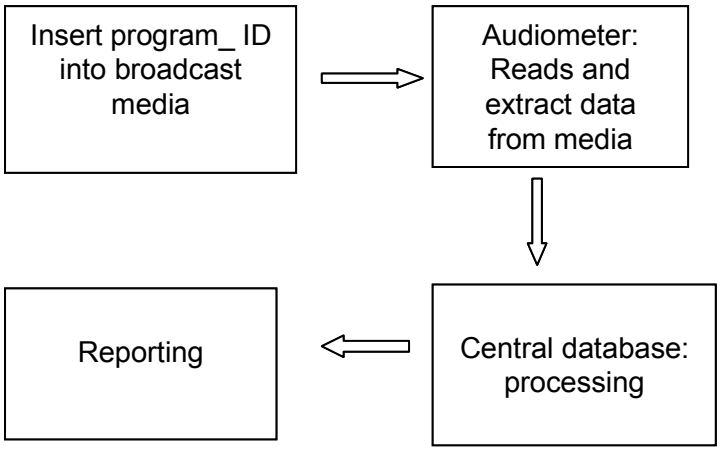

Figure: Proposed audience measurement by watermarking.

The advantages of using digital watermarking are that it can be applied in radio, TV and internet audience measurement with less complexity in terms of hardware requirement. If blind embedding is used then the Program_ID can be extracted without reference to the original insertion at broadcast time. It can also allow for other side information embedding useful for program identification.

Conclusion / Future

Most African countries with the exception of South Africa have developed a model for audience measurement system. With the interaction of connected consumer electronics and digital media there is the creation of vast and limitless amounts of user data, now commonly coined "Big Data." Given benefits and potential in the advertising industry, government, educational and scientific application stakeholders need to develop smart ways to manage the Big Data in their audience metrics and other aspects of the economy. The development of a unified audience metric standard is a difficult but important goal, and stakeholders should strive to provide full transparency to users with respect to tracking data as well as opt-out mechanisms. The future work will be on the integration of audience measurement and big data into Nigeria educational system and E-governmental purposes.

References

Arab News. (2014, 5 April). *Lectures get a glimpse on big data technologies*.

Audience measurement. (n.d.). Digital Watermarking Alliance. Retrieved from http://www.digitalwatermarkingalliance.org/app_audience.asp

Bornman, E. (2008). *Measuring Media Audiences*, chapter 12, pp.1-52.

Coffey, S. (2001). Internet audience measurement: A practitioner's view. *Journal of Interactive Advertising, 1*(2), 10-17.

Hoover, J. N. (2012). Government's 10 most powerful supercomputers. *Information Week*.

How big data analysis help President Obama defeat Romney in 2012 Elections. (2013). Bismol Social Media news.

Kalil, T. (2012). *Big data is a big deal*. The White House.

Katzenbeisser, S., & Petitcolas, F. A. P. (2000). *Information hiding: Techniques for steganography and digital watermarking*. Norwood, MA: Artech House Books.

Lim, C., Choi, J. H., Nam, S. W., & Chang, J. H. (2013). A new television audience measurement framework using smart devices. *Multimedia Tools and Applications*, 1-20. DOI 10.1007/s11042-013-1658-7

Megías, D., Serra-Ruiz, J., & Fallahpour, M. (2010). Efficient self-synchronised blind audio watermarking system based on time domain and FFT amplitude modification. *Signal Processing*, *90*(12), 3078-3092.

Mooney, A., & Keating, J. G. (2005, June). Generation and detection of watermarks derived from chaotic functions. In *OPTO-Ireland* (pp. 58-69). International Society for Optics and Photonics.

Muller, S. (2001). *Audience measurement*. Elsevier Science Ltd.

Petitcolas, F.A.P., Anderson, R. J., & Kuhn, M. G. (1999). Information hiding - A survey. *Proceedings of the IEEE*, 87(7), 1062–1078.

SAS Institute. (2014). *Big Data: What is it and why it matters*.

Seitz, J. (Ed.). (2005). *Digital watermarking for digital media*. IGI Global.

Sloan Digital Sky Survey. (n.d.) In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Sloan_Digital_Sky_Survey

Way, H. (2014). Managing big data: Audience measurement and ad targeting. *E-Commerce Times*. Retrieved from http://www.ecommercetimes.com/story/79902.html

Biographies

Agbaje M.O. is a lecturer and currently working on his Ph.D at Babcock University, Nigeria. His research interests are Information security, Digital watermarking, A.I and Embedded systems.

Oludele Awodele has a Ph.D in Computer Science. He is currently the H.O.D of Computer Science department, Babcock University. His areas of interest are A.I and Computer Architecture. He has published scientific articles in several journals of international repute.

A.C. Ogbonna Ph.D is presently the dean of School of Computer Science and Engineering, Babcock University ilisan Remo Ogun State, Nigeria. He can be contacted at acogbonna06@yahoo.com