Afr. J. Lib. Arch. & Inf. Sc. Vol. 26, No. 2 (October 2016) 93-96

Editorial Feature

Big Data Industry: Implication for the Library and Information Sciences

Stephen Mutula

Information Studies Programme
University of KwaZulu-Natal
Pietermaritzburg, South Africa
Mutulas@ukzn.ac.za

University of Pittsburgh (2007) defines Big Data as the sets of data that are so large and complex to use effectively and efficiently. Chen and Zhang (2014) on their part define Big Data as a collection of very huge data sets with great diversity of types, that it is difficult to process by using state of the art processing approaches or traditional data processing platforms. They point out that a data set can be called Big Data if it's formidable to capture, curate, analyse and visualise using current [or conventional] technologies. Penn State College of Information Science and Technology (nd) looks at Big Data differently as a process that is concerned with the exploration, development, and applications of scalable algorithms, infrastructures, and tools for organising, integrating, retrieving, analysing, and visualising, large, complex, and heterogeneous data. SAS Institute (nd) characterises Big Data in five ways which can deciphered in 4Vs and C, namely: Volume (that organisations collect data from variety of sources including business transactions, social media, sensor or machine to machine data); Velocity (that data streams in at unprecedented speed and must be dealt with in timely manner); Variety (that Big Data comes in all types of formats-structured, numeric, unstructured text documents, email, video, audio, and more); Variability (that Big Data flows can be highly inconsistent); and Complexity (that Big Data comes from multiple sources, which make it difficult to link, match, cleanse and transform across systems).

Realistically, Big Data cannot be measured in Megabyte (for example, a book or photo) or Gigabyte (for example, a movie), but in Terabyte (for example, all books in the world), Petabyte or Exabyte (for example, all books in multimedia formats in the world), Zettabyte or Yottabyte (for example everything recorded in human history) (Gray and Belew nd) because of huge volumes involved. Bieraugel (2016) adds that Big Data cannot be stored or analysed by conventional hardware and software because traditional hardware can handle Megabyte and Kilobyte sized data sets, while Big Data can handle Terabyte and Petabyte sized data sets. Hagstroem (2015) and ; Shaw(2014) point out that the sources of Big Data include Web services, social media, open data services (such as governments), archived information in libraries, repositories, digital archives, phones, credit cards, television, computers, the infrastructure of cities, sensor equipped buildings, trains, buses, planes, bridges, factories… and [surveillance systems], satellite, navigation systems and more. Big Data was estimated at 2.8 Zettabytes of the global data in 2012. By 2020, the data is expected to increase by 50 fold. Hagstroem (2015) points out that only 0.5 percent of available data globally are currently analysed because of the limitations of traditional computational, data storage and retrieval tools.

Big Data analytics has therefore emerged as a field of data science or e-science that provides high performance computational tools to analyse the huge volumes of data (Big Data) generated daily

worldwide in order to afford insights into untapped and trapped data in the traditional relational database systems of the past (Hagstroem, 2015). Big Data analytics has therefore great potential to generate answers to myriad of complex problems facing humanity from climate change, conflicts, biodiversity, earth tremors, poverty, hunger, migrations, and insecurity among others by illuminating the hidden patterns into huge volumes of data stored in databases and data warehouses that remain untapped.

The importance of Big Data need not be over emphasised. Chen and Zhang (2014) assert that Gartner listed the top ten strategic technology trends for 2013 and top 10 critical technology trends for the next five years, and Big Data is listed in both the two. Shaw (2014) in an article in the Harvard Magazine of March-April 2014 titled 'Why Big Data is a big deal' points out that understanding Big Data leads to insights, efficiencies, and saved lives. SAS Institute (nd) observes that Big Data can be analysed from any perspective to find answers and enable cost reductions; time reductions; new product development; strategic business moves; optimised offerings and smart decision making. Moreover, by analysing Big Data, businesses can be helped to determine root causes of failures, defects and issues, and detect fraudulent behaviour in an organisation. Besides, Big Data is being widely used in banking to understand customers and boost satisfaction; in education to identify at-risk students, enable students to make adequate progress, implement effective system for evaluation and support; in government to manage utilities, running agencies, dealing with traffic congestion or preventing crime; in health care to manage patient records, prepare treatment plans, generate prescription; in manufacturing to improve production and quality; and in retail to build customer relationships and design market differentiation programmes (SAS Institute nd). Davenport and Dyche (2013) assert that…the primary value from Big Data comes not from data in its raw form, but from the processing and analysis of it and the insights, products, and services that emerge from the analysis…

Bieraugel (2016) attributes the increasing growth and interest in the Big Data to the motivation to lower costs of servers to house the data, the release of open source software tools to manage distributed computing, the creation of massive data sets, and the need for businesses and other entities to leverage value out of the data they collect. SAS Institute (nd) adds that faster processors, distributed Big Data platforms (for example, Hadoop), parallel processing, virtualisation, large grid environments, high connectivity and throughputs are other factors driving the growth of Big Data industry.

Chen and Zhang (2014) are of the view that Big Data has the potential to make prominent growth of the world economy by enhancing the productivity and competitiveness of enterprises and the public administrations. In the US, for example, it is estimated that Big Data industry produces 140,000 to 190,000 deep analytical talent positions and 1.5 million data savvy managers. Big Data science is being used in nuclear research, astronomy, e-commerce, atmospheric science, genomics, biogeochemistry, bioinformatics, social network analysis, and retail entities.

A number of academic disciplines are leveraging Big Data. For example, according to Shaw (2014) government faculties in the US are doing some type of data analysis with scholars in sociology, economics, public health and medicine. In marketing, Big Data is being used for client differentiation and preferences based on purchase analysis or trends. In law, Big Data is being applied to predict the likely outcome of cases, especially in supreme courts. Credit card companies are using Big Data to understand and evaluate the risk of default, and crime fighting agencies are using Big Data to allocate resources by predicting when and where crimes are likely to occur. In the United States, Big Data has been used to predict flu outbreaks faster than is possible using patient admission records.

Big Data is therefore having transformative impact in all academic fields. In the library and information sciences, though Chen and Zhang (2014) assert that Big Data has drawn huge attention from researchers, major strides have not been made. Potential areas of use of Big Data in library and the information sciences could include: search strategies development for Internet searching, research data management, data curation, accessible internet text indexing, natural language processing and retrieval, data privacy, access and usability issues; developing user friendly interfaces for data mining; keyword indexing, Boolean searching, relevance feedback, recall and precision enhancement, advanced information retrieval processes, and information extraction ,among. Bierraugel (2016) points out that

it is important for librarians to know and understand Big Data and how it can be used to facilitate basic research, how companies leverage Big Data, how Big Data creates competitive advantage for organisations, and how they (librarians) can make Big Data visible, accessible and usable by creating taxonomies, designing metadata and developing systematic retrieval methods. In addition, librarians can use Big Data tools to analyse data sets to make them simple, searchable and usable.

In the United States, some LIS schools have pioneered curricular and research into Big Data domain. For example, the University of Pittsburgh has introduced courses that are aimed at developing skills for employment in the Big Data industry. Such courses include, but are not limited to data mining, adaptive information systems, cloud computing, data analytics, information visualisation, and neural networks. Penn State College of Information Science and Technology (nd) on the other hand offers information retrieval and search, scalable machine learning, learning predictive models, semantic complex event processing (CEP), Big Data privacy and security, discovery Informatics, and Big Data applications in informatics (including Health Informatics, Security Informatics, Social Informatics), among others. Cornell University (nd) offers in contrast the following courses in Big Data domain: data science, industrial data and systems analysis, data mining and machine learning, ubiquitous computing, natural language processing, and designing technology for social impact. Similarly, Harvard has introduced a course in data science, and is contemplating offering other new courses in Big Data domain such as computation biology, and quantitative genetics in order to leverage improved methods of processing and mining Big Data.

Analysis and use of Big Data is not without challenges. University of Pittsburgh (2007) asserts that the major challenge associated with Big Data is the rate of growth, diversity of multiple data sets and formats. The other challenges occur in data capture, storage, searching, sharing, data inconsistencies and incompleteness, scalability, timeliness and data security because of variety of data formats of the data from different sources (Chen and Zhang, 2014).

The emerging field of Big Data has therefore tremendous impact in all academic fields and promises great potential in creating new skills in various academic sectors including the information sciences in the areas of data management, curation and archiving, search and retrieval, interdisciplinary research, and the LIS curriculum. The other potential areas to grow skills in the library and information science are high intensity performance computing, advanced statistical and computational methods, virtual reality systems, diversity formats data management, digital preservation and curation, among others. Big Data provides another milestone in the development of science for the librarians to reinvent themselves and become more relevant in a dynamic and rapidly changing information environment that has attracted many players.

References

Bieraugel, M. (2016). Keeping with…Big Data. Available at: http://www.ala.org/acrl/publications/keeping_up_with/big_data. [Accessed 10 September 2016].

Chen, CLP and Zhang, C Y. (2014). Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. Information Sciences 275 (10 August 2014), 314-347.

Cornell University (nd). Data Science. Available at: http://infosci.cornell.edu/academics/undergraduate/undergraduate-concentrations/concentrations/data-science. [Accessed 10 September 2016]

Davenport, T. H. and Dyche, J. D. (2013). Big Data in Big Companies. Available at: https://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf [Accessed 10 September 2016]

Gray, R. and Belew, R. (nd). What is information retrieval? [Slide Presentation]. Available at: http://slideplayer.com/slide/5835585/ [Accessed 10 September 2016].

Hagstroem, M. (20015). Big Data Analytics for Inclusive Growth: How Technology Can Help Elevate the Human Condition. Available at: http://reports.weforum.org/global-information-technology-report-2015/1-8-big-data-analytics-for-inclusive-growth-how-technology-can-help-elevate-the-human-condition/ [Accessed 10

September 2016].

Penn State College Information Science and Analytics, Modeling and Informatics. Available at: https://ist.psu.edu/node/1633 [Accessed 10 September 2016].

SAS Institute (nd). Big Data: What is it and why it Matters. [Online]. Available at: http://www.sas.com/en_us/insights/big-bid-data/what-is-big-data.html [Accessed 10 September 2016].

Shaw, J (March-April 2014). Why big data is a big deal. Harvard Magazine (March-April 2014). Available at http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal. [Accessed 10 September 2016].

University of Pittsburgh (2007). Big Data Analytics Specialisation. Available at: http://www. is chool.pitt.edu/ist/degrees/specializations/big-data.php [Accessed 10 September 2016].