The Evolution of Data. From Data to Big Data.
Are we ready for the big data technology in the library community?

Rania Ramadan Osman
Head of Libraries of the Future Section
Bibliotheca Alexandrina, Alexandria, Egypt
rania.osman@bibalex.org

1 ABSTRACT

For many years, libraries have been the knowledge centres and the places where you can find large amounts of data in different formats both physical and electronic. The libraries were mainly for researchers, but with the evolution of information and rapid progress of the technology that affected the library community in many aspects, libraries have extended their role to become more and more involved in the communities they serve. The evolution affected the library itself, yet the librarians, especially in the Arab region, are still behind these technologies, trying hard to cope and survive. There is a gap between the rapid progress of the technology and the qualifications of the librarians who should adopt the technologies. The data becomes so large and the formats the librarians were familiar with changed, even the librarians’ tools transformed from printed to more complicated formats. Recently, new terms were introduced to the library community, linked data and lately big data. Actually, these are not new terms, they are very popular in the information technology field, but for the librarians they are fresh and still not easy to comprehend. The research presents the main challenge that usually prevents Arab librarians from interacting with new technology; language barrier. We need translations and Arabic equivalents for the terms that occurred recently and were mainly adopted from other disciplines.

The purpose of this paper is to present the actual position of the Arab libraries and librarians in the light of all these new applications. Are we as Arab librarians aware of the usability of the big data in the libraries; this research paper depends mainly on the results of a survey targeting the information service specialists in the public libraries in Egypt in an attempt to answer five main questions about their knowledge about big data.

2 INTRODUCTION

Modern libraries of today should place themselves as community anchors to provide the enormous learning possibilities, where the librarians can provide patrons with all the new findings of technology.

Therefore, librarians need to know the basics of technologies and the crucial issues that the community of scholars and business men talk about recently. One of the main topics is big data and how it affects academic research. Librarians need to know how firms benefit from big data applications, how such data mining provides a competitive advantage, and how students might need to deal with big data sets in
future. They need to know the impact of big data on the library community, how we could make full use of library data, and how can we transfer library data to big data. They should be aware that big data is becoming more common in their work no matter in which area they are specialized. Librarians on the era of technology should be facilitators of research process not only providers of information for researchers, We as librarians need to be part of the technology and not isolated in our achieves working on old applications, so if the future directs us to big data, we are required to know how big data is to be used in libraries.

3 RESEARCH QUESTION

AnnaLee Saxenian presented her expectation to the characteristics of the data librarian to be as follows “A data librarian has a special set of responsibilities around stewardship and curation. . . defining standards, storing data . . . and organizing data in a way that makes it more accessible. And it may be a bit of an uphill battle.” In the Arab region we still think of big data as a theoretical concept not a trend that we aspire. A needed move and transition in our community that we should be part of it. It’s about time to move from collection based to community orientated services, we have already moved from physical to digital formats in the last years and we are now using the digitization techniques as one of the main trends to preserve our heritage and collections. When we mention the digital age and ways of preservation we need to make sure that librarians are aware of the data processing and data analysing systems that would empower the libraries to move ahead and again this kind of transition would never be maintained deprived of high competency librarians. This research paper introduces one main question that we all think about when presenting any new technologies, Are the daily users of this techniques need it and if yes are they qualified to deal with such technologies in the time being.

4 RESEARCH SIGNIFICANCE

Fei Hu stated that ”The growing gap between data and users calls for innovative tools that address the challenges faced by big data volume, velocity and variety”. (Hu, Fei, 2016). This means that Big Data is not a new concept, yet it is new to the library community. Libraries deal with various huge amounts of data like the cataloguing tools, subject thesauri, catalogues, and metadata records, could we consider all these sorts of data sets as big data. Most Organisations now are trying to understand big data concepts to analyse and categorize the types of data they have and to make better use of their data. Moreover, there is a great opportunity to save money by better understanding of big data usage, so libraries could benefit from this to deduct their expenses. This research paper tries to shed light on the importance of provide special training programs by scientists and information technology professionals to introduce and explain the basics of technology trends to the librarians on how to deal with technology updates like big data. Moreover, the research presents the challenges the librarians face when dealing with any new topic

1 http://publiclibrariesonline.org/2015/05/data-librarians-in-public-libraries/
associated with technology or affiliated to the Information technology discipline. It is a great barrier for any librarian to deal with all these concepts, they are really complicated and most of the resources in English which make it very difficult to many librarians to comprehend the text or even to convey the message of the author.

5 RESEARCH GOALS

- Decreasing the gap between Arab librarians and the new technology by introducing all the new concepts and implementations
- Benefit from big data concept in library business and how to transform library data into big data
- Initiative to teach big data as part of the knowledge organization syllabus in school libraries in the Arab region
- Providing Arabic equivalent terms to the new concepts like big data and smart data in an attempt to support the Arabic readers to be aware of these issues in their mother tongue

6 RESEARCH METHODOLOGY

The methodology used for gathering data was mainly questionnaire. The design for the questionnaire was based on the research question of the study. It was a closed ended question and respondents were asked to tick only the most appropriate response. Where no options were designed they were asked to provide answers. A total of 100 questionnaires were mailed to the five public libraries in Egypt as it was easier for the author to select libraries from his country rather than sending the questionnaire to many libraries in the Mena region. The author asked the respondents to mention their job title and to briefly define their job description and years of experience. A total of 43 responses were received as shown in the figure 1 below.

Figure 2: Questionnaire Results

2 Figure 2: Questionnaire Results

2 https://bigdatainarabic.wordpress.com/
The questionnaire entailed 4 questions and it was divided into 2 sections. The respondents were mainly information services specialists who work closely with the patrons.

The outcomes from the questionnaire were not as expected; the author received 15 responses out of 43 whom managed to understand the questions and answers them properly. 12 respondents failed to comprehend the questions and they provide inappropriate answers. 16 respondents did not complete the two sections and they mention that they are not aware of the topic big data.

The questionnaire presents the case in Egypt that has more than 350 public libraries, 285 specialized libraries and 519 academic libraries. According to the results of the questionnaire, we face a real problem in the library community with the literacy of the librarians and the language barrier is still the main challenge, most of the participants failed to provide proper answers because they did not understand the questions written in the English language. Most of the respondents had more than five years of experience and 25 of the respondents tick yes when they were asked if this was their first time to hear about big data term. The good thing about the feedback of the respondents is that they were all so eager to learn and know more about big data and its usage in libraries. The most common questions that I received after sending the questioners back was if there is any opportunity to have more workshops or references in Arabic that could give them more clues about the topic. I tried to collect some of the definitions and characteristics of big data from different sources in the light of the shortage of the literature written that target the libraries and library data. So there was a great confusion between the different concepts; big data and Library data. In the light of the finding of the questionnaire, the next section of the paper will present the characteristic of library data and bi data as well.

7 CHARACTERISTICS OF LIBRARY DATA

Librarians used to dear with different sorts of data in different categories, yet there were many trials to organize these sorts of data to facilitate the daily work of the librarians. The data describing the resources were mainly descriptive tools like AACR and recently the new standard RDA. Data identifying the subject of the resources and mainly called subject headings that are extracted from different Thesauri like Library of Congress Subject Headings and many others. Data that classify the subject under certain discipline of knowledge with reference to other specification that has relation with the geographic places and ethnic groups; they data are mainly classification tools like Dewey Decimal Classification either full edition or abridged version or even web version. In library community, the creation of these types of data is seen as the main role of librarians and information specialist; these are mainly the sorts of data they are familiar with. (Fredrick, D., 2017)

The problem with library data was always the complexity of the way of registering the data and the indexing tools used to retrieve the data. Also, The huge amount of the library data recorded by the librarians with great efforts in an attempt to maintain the higher quality standards, yet most of the data is contained in the library databases and catalogues not available via search engines like Google. There are some initiatives to replace the old tools with new web based ones that depend mainly
on the fundamentals of semantic web and linked data like BIBFRAME which will replace MARC 21 formats in order to be able to break all the bounties of library data and release the data from the catalogues.

**8 THE DEFINITIONS AND CHARACTERISTICS OF THE TERM “BIG DATA”**

Big data is a one of the terms that emerged recently and most of people became familiar with the term, yet not everyone in the library community is aware of its usage and applications in the library community. Big data is an abstract term and there are various definitions for many researchers, mainly from the information Technology field and from others fields, just few from the library and information science as it is still a new-born term for the library community and we attempt to muddle through with its concepts and characteristics.

According to IBM the world creates 2.5 quintillion bytes of data. For the record, quintillion has 18 zeros. IBM contends that 9 percent of the data published today has been created in the last two years alone. (Matt, et al., 2014)

Big data is the result of the rapid growth of the internet as described by Iafrate (Iafrate, F., 2015)

According to lafrate, big data is associated with data processing (Iafrate, F., 2015) while Pawr referred to big data as large volume data that can't be mined, collected, or even managed by traditional analysis implements. (Pawr, A.M., 2016). On the other side Mckingsly mentioned that big data is used to refer to data sets whose size is large and exceed the capability of existing data analytics tools to mine, collect, store, process and analyze within a specific amount of time (James et al., 2011). Chen and Zhang pointed out to big data as a bundle of large data sets that is gathered from various sources with numerous formats therefore it is difficult to be managed by traditional data analytical systems to maintain the big data processing and analysis. (Chen and Zhang, 2014)

De Mauro defined Big data as “the information asset characterized by such a high Volume, Velocity, and variety to require specific Technology and an Analytical Methods for its transformation into Value” (De Mauro, et al., 2016), while Iafrate agreed with the last definition and referred to big data as “the data that is principally characterized by the four “V”s. They are Volume, Variety, Velocity, and Value”, yet Iafrate added the value to the characteristics mentioned before and this is mainly associated with smart data rather than big data. (Iafrate, F., 2015)

Based on the previous definitions, big data has certain characteristics mainly Volume, Velocity, and Variety and they are known as 3V’s as shown in figure 2. Volume refers to the large size of the data that grows rapidly growing. Velocity refers to the speed of the data generation as stated by Chean that the data collection, data mining and data analysis should be managed quickly in order to make full use of the probable significance of these data While Variety refers to the various formats either structured or unstructured data obtained from various sources. (Chen and Zhang, 2014). Some researchers added the fourth V which to represent the Value as one of the main characteristics of the big data as noted. (Pawr, 2016) Value refers to the importance and significance benefits to the users and when using the big data in their business. (Chen and Zhang, 2014)
IMPORTANCE AND CHALLENGES OF BIG DATA TO LIBRARIES

Big data is like any new technology that could be so beneficial, yet it has more challenges and will bring up issues that you need to solve to reach the level of importance you aspire. The most important significance of big data for libraries and information organizations is how to facilitate the analysing the big sets of data to extract the essential needs of the users (Provost, F. & Fawcett, T., 2013). Also, the current data management and analysis systems used in libraries are designed to be applied to structured data while big data processing technology presents the privilege of analysing both structured and unstructured data (Kitchin, R. & McArdle, G., 2016).

Big data also could empower decision making in libraries as stated by Gartner that more than 75 percent of companies are investing or planning to invest in big data in the coming years due to the importance of data analysis to support making proactive decisions (Gartner, 2015). Big data processing of data could help Libraries to analysis the worthy and unworthy data they have and to make it easier for data driven decisions. Decisions related collection development, spaces’ choices, researchers needs, and even data extracted from users’ reviews on the social media and even what’s app messages and alerts to be sent regularly to the users. Also, questions and surveys and events surveys that are emails to the users could also be analyzed by big data applications and enable the librarians and deskin makers in libraries to act according to the accurate results gathered and analyzed properly.

These are some more aspects to big data usages in libraries, but our main focus in this research paper is to present the main challenges of applying big data in libraries and the challenges that the librarians face to cope with big data trends.

3 Figure 1: Big data characterises

3 https://bigdatainarabic.wordpress.com/
10 CHALLENGES OF APPLYING BIG DATA IN LIBRARIES

The challenges of applying big data in libraries essentially comprise data accuracy, data reduction and compression, data confidentiality and security and big date processing system and technologies. To apply big data in libraries we need to enrich the library database and enhance the skills of librarians. Big data application in libraries depend mainly on specific aspects: data accuracy, data reduction, big data processing system and technology, and data confidentiality and security. (Li, Jun, et al, 2017)

By data accuracy we mean the unstructured data that the library deal with, to enable the application of big data we need to organize the structure of the unstructured and semi structured data to facilitate the analysing processing and emphasizing on the accuracy of the data provided. As for the data reduction, the libraries should filtered the data into two types the worthy and unworthy data and evaluate the data in order to be easier for transition into big data sets.

In the previous section, the researcher mentioned the privilege of using the big data processing management system, yet the challenges with this will be the high cost that may prevent many libraries of applying big data. The last challenge is the data confidentiality as we mentioned earlier that due to the lack of experience and technology skills, the librarians would not be able to deal with big data by themselves and libraries would tend to hire some professional data processing agencies and this could result in making the data of the libraries exposed and vulnerable to others.

11 CONCLUSION

This research paper tried to present one of the key and central topic that attract many of the researchers recently and we all as professional librarians and researchers are trying to map with all the trends that emerging trends of technologies bearing in mind the different categories of qualifications of librarians and their responding to such technologies. The paper introduced the characteristics of both library data and big data in an attempt to answer the research question Are we ready for big data. Still the answer to this question needs more research and more cooperation between the data specialists and the librarians to prepare guides and workshops and special trainings to qualify the librarians to be able to deal with big data in the future. In other words, the main obstacles that could prevent libraries from applying big data application framework is the manpower and lack of information technology skills.

12 REFERENCES


