THE USE OF LEAN SIX SIGMA METHODOLOGY IN DIGITAL CURATION

G. Arcidiacono[1], E. W. De Luca[1], F. Fallucchi[1], A. Pieroni[1]
[1]*Department of Innovation & Information Engineering (DIIE)*
Guglielmo Marconi University, Via Plinio 44, Rome 00193, Italy

Abstract.
In this paper, we give an overview about the current research in Big Data and Digital Curation with a focus on Lean Six Sigma and discuss how this methodology can help the Digital Curation lifecycle. For instance, the application of the Lean Six Sigma methodology is presented and discussed with a special focus on the selection, preservation, maintenance, collection and archiving of digital information, the so-called Big Cultural Data. The aim of our work is to present a methodology for Digital Curation lifecycle, asserting that all the actions belonging to the Data Curation may be performed and optimized by using DMAIC (Define, Measure, Analyze, Improve, Control) phases of Lean Six Sigma.

Keywords: Lean Six Sigma, Big Data, Digital Curation, Digital Humanities

1 Introduction

The vast amount of data in the field of cultural heritage makes it often difficult for the interested person to retrieve the desired information. It has been thus imperative need to introduce automatic methods that increase the relevance of hits by semantic search. This can be achieved with the use of the Semantic Web. However, the various meta-languages that have already been adopted in the Semantic Web are usually not compatible with each other.

Furthermore, Digital Curation is generally referring the process of establishing and developing long-term repositories of digital assets for research issues. Enterprises are starting to utilize Digital Curation to improve the quality of information and data within their operational and strategic processes.

One of the biggest challenges is to extract the relevant information from the huge amount of data available in the digital world. In this paper, we give an overview about the current research in Big Data and Digital Curation with a focus on Lean Six Sigma (LSS) and discuss how this methodology can help the Digital Curation lifecycle.

2 Literature Review and research scope

The connection between Lean Six Sigma (LSS) and Big Data has been increasingly established by the contribution of LSS to accelerate the process of extracting key insights from Big Data, while highlighting how Big Data can bring new light and innovation to projects requiring the use of Lean Six Sigma (Fogarty 2015). Big Data have been given different definitions (Franks, 2012; Dumbill, 2012; Gobble, 2013); however, it is commonly agreed that Big Data is often defined as an increased amount of data thanks to the internet, and to the data related to the use of wireless devices, leading to the opinion that Big Data are the next step in innovation (Gobble, 2013).

While the connection between LSS, Big Data and Digital Humanities Research is still under preliminary debate, the connection between LSS, Big Data and manufacturing has been established and discussed in previous research. In fact, as regards manufacturing systems, the influence of the onset of Big Data and Big Data analytics has been well identified, as well as the relationship of IoT (Bi & Cochran 2014; Bi, Xu, & Wang 2014) and Big Data analytics (Mo & Li 2015), which facilitated information visibility, and increased the level of automation in design and manufacturing engineering (Wang & Alexander 2015). However, Parker (2014) commented that only about 10% of the value potential of the information collected

was actually utilized to enhance the level of management productivity. This because high volume manufacturing system architecture may be different than low volume products and there may be great difference in initial system maturity (Bi 2011).

2.1 Metadata, Digital Curation and Digital Humanities

Metadata is used to describe objects or processes regardless of the domain. There exist a large number of metadata standards and formats. The cultural heritage domain deals with developments like EAD (Encoded Archival Description), EAC-CPF (Encodes Archival Context - Corporate Bodies, Persons, and Families), CIDOC-CRM (CIDOC Conceptual Reference Model) or METS (Metadata Encoding and Transmission Standard) to describe specific objects. EAD is an XML-based standard for representing the structure of archival finding aids. Developed in 1993 it is based on ISAD(G) (General International Standard Archival Description). EAD is widely used in the USA. In Europe it is implemented more and more. It is expected, that most of the data collected by the project will be in EAD. EAC-CPF was developed as a supplement to EAD and was introduced in 2008. The description of persons, families and institutions that are associated with the creation, preservation or use of the archive or in any other way is in the focus of this standard. CIDOC-CRM is the acronym for International Documentation Committee of the International Council of Museums - Conceptual Reference Model. The data model has the goal of data sharing and data integration of heterogeneous data sets between different systems and disciplines of the cultural heritage sector, primarily museums but also archives. Semantic definitions are proposed for the transformation of distributed information into comprehensive global resources. METS is the XML-based description of descriptive, administrative and structural metadata for a digital collection. Here, the digital objects are in the foreground.

To represent and use the metadata provided by different institutions in different metadata formats a top level exchange format is needed. Additionally, we need automatic processes to support data curation. For using such automatic processes in Digital Humanities, we need to first understand or give the definition of ‘Digital Humanities’. This research field is a building bridge between information sciences and the various humanities disciplines.

However, openness was always associated with a need for introspection and some tentative boundaries definitions (Risam 2015). This research domain is defined dynamically in the negotiation of these tensions as discussed by several Digital Humanities scholars (Unsworth 2002; Svensson 2009; Rockwell 2011). In this paper, we understand Digital Humanities as a scenario for the Lean Six Sigma and its use with Big Data.

2.2 Lean Six Sigma and Big Data

Big Data processing can successfully be integrated by Lean Six Sigma (LSS) can, as LSS is a strategy addressing entire process systems, aimed at reducing non-value-adding activities, this by processing a huge amount of data. Ideally LSS can be implemented to optimize performances of a varied range of systems, from the least to the most complex, even when limited resources are available, to allow those same limited resources to be spent most productively, such as within hospitals (Arcidiacono, Wang, & Yang 2015). This because LSS isolates the main critical stages and features of the whole process in sub-phases: problems are deconstructed into smaller areas, to make process knowledge more accessible, and to solve process issues with surgical precise actions (Arcidiacono, Costantino, & Yang, 2016).

Axiomatic Design (AD) is the tool to design the LSS training and process management model, because it provides relevant criteria to critically analyse design.

Most significantly within the scope of this research, AD is indeed a flexible tool suitable to be effectively applied to varied range of context and scenarios whereby process improvement and optimization is required (Arcidiacono, Giorgetti, & Pugliese, 2015; Arcidiacono & Placidoli, 2015). The DMAIC (Define, Measure, Analyze, Improve, Control) could be applied as the methodology framing the entire

optimization process, to determine the dependence of a system and process reliability (Arcidiacono & Bucciarelli, 2016). Therefore, LSS could be effectively implemented in Digital Humanities Research, specifically within Digital Curation, to isolate relevant data, avoid data obsolescence, and enhance data availability and high-quality research, by optimizing data extraction and its related functions.

3 Our Approach: the use of LSS Methodology in Digital Curation

The exploration and the definition of the context and boundaries that belong to the Big Data Digital Humanities Research area continue to be considered as an unsolved and complex system. There are studies in literature (Kaplan, 2015) that attempt to represent Big Data Research in Digital Humanities as a structured field, by proposing a division of three concentric areas of study: Big Cultural Data, Digital Culture and Digital Experiences. The aim of the author consists of proving that this huge amount of information can be organized as structured field and, consequently, can be characterized by common methodologies.

The goal of this paper, instead, consists of investigating the application of well-known methodologies of process improvement and process optimization, such as Lean Six Sigma methodology, to the Digital Curation aspects of the Digital Humanities, where Digital Curation consists of selection, preservation, maintenance, collection and archiving of digital information, with particular focus on the so-called Big Cultural Data. In other words, Digital Curation involves maintaining, preserving and adding value to digital data throughout its lifecycle. The active management of data reduces risks to their long-term value and mitigates the threat of digital obsolescence. Meanwhile, curated data in digital repositories may be shared among a wider research community, increasing the intrinsic value of the Cultural Data, as shown in Figure 1.

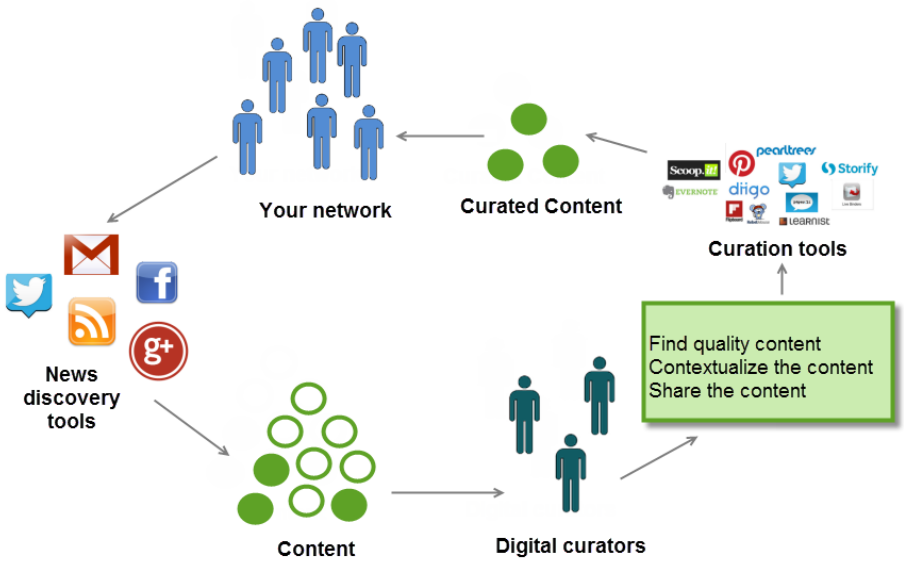

Fig.1: Digital Curation rules[1]

Furthermore, Data Curation enhances the long-term value of existing data by making it available for further high quality research. On the other hand, Cultural humanists, involved in the Digital Curation aspects, are increasingly engaged with curating and making accessible the digital materials.

As said above, Lean Six Sigma methodology is successfully and widely used in many areas such as government, industry, healthcare, and education. The

1 The source of Fig.1 is the blog article "Digital Curation: putting the pieces together" by Sue Waters, available at http://suewaters.com/author/suewaters/. [Last accessed 10th November, 2016]

methodology is based on the use of the DMAIC approach, a data-driven quality strategy, as an instrument usable during the phase of extraction, analysis and sorting of the data (Fogarty, 2015). DMAIC is an acronym representing the five phases that make up the optimization process. In particular, *Define* the problem, the improvement activity, the opportunity for improvement, the project goals, and the customer (internal and external) requirements. *Measure* process performance. *Analyze* the process to determine root causes of variation and poor performance (defects). *Improve* process performance by addressing and eliminating the root causes. *Control* the improved process and future process performance.

The processes of extraction, analysis and sorting the data allow to predict the future trends and to achieve advantages in all environments. Lean Six Sigma is a complex methodology where the accurate organization is able to observe and mitigate the errors and the deviations occurring in its operations by applying strict rules. This paper aims at initiating a new line of research that consists in investigating how methodologies, such as LSS, may automate and optimize functions, such as categorization, classification, clustering and digitalizing Big Cultural Data. This new line of research should study, first of all, the meaning of Data Curation, which are the actions to be performed and which actions may be automated and supported by digital instruments, such as LSS. The first action, as shown in Figure 2, consists in describing and representing the information: appropriate standards should be used in order to describe metadata, so that it can be controlled over the long term.

Furthermore, all metadata and associated digital material should be represented in appropriate formats. The second action consists in building a preservation strategy: this action is important to plan for preservation throughout the data lifecycle.

Collaborating, supervising, and participating are the actions to be performed in order to supervise data creation activities and to assist in the creation of the standards to be used. Finally, curating and preserving represent the actions to be performed to take into account the managerial and administrative aspects.

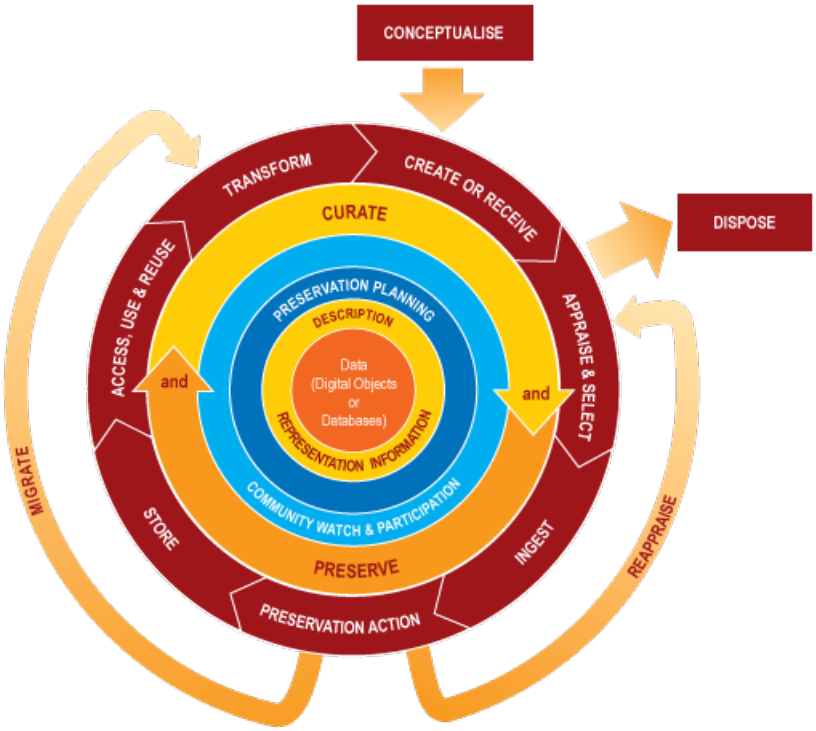

Fig.2: Actions of Data Curation Lifecycle[2]

2 http://oxdrrc.blogspot.it/2008/12/research-data-management-services.html "Research Data Management and Curation Services Framework", Oxford University Digital Repository

4 Final Remarks and Conclusions

In this paper we have given a research overview on Big Data and Digital Curation and we have proposed the application of well-known methodologies of process improvement and process optimization, such as LSS methodology, to the Digital Curation aspects of the Digital Humanities. Thus, the LSS is applied for the selection, preservation, maintenance, collection and archiving of digital information, with particular focus on the so-called Big Cultural Data. Furthermore, we have discussed how this methodology may help the Digital Curation lifecycle, asserting that all the actions belonging to the Data Curation may be performed and optimized by using DMAIC phases of LSS.

5 References

Arcidiacono, G., & Bucciarelli, L. (2016). TRIZ: Engineering Methodologies to Improve the Process Reliability. Quality and Reliability Engineering International Journal, 32 (7): 2537-2547.

Arcidiacono, G., Costantino, N., & Yang, K. (2016). The AMSE Lean Six Sigma Governance Model. International Journal of Lean Six Sigma, 7 (3): 233-266.

Arcidiacono, G., Giorgetti, A., & Pugliese, M. (2015). Axiomatic Design to improve PRM airport assistance. In: Proceedings of ICAD 2015, 9th International Conference on Axiomatic Design, edited by M. K. Thompson, D. Matt, A. Giorgetti, N. P. Suh, and P. Citti P., 106-111. Red Hook: Curran Associates.

Arcidiacono, G., & Placidoli, P. (2015). Reality and illusion in Virtual Studios: Axiomatic Design applied to television recording, In: Proceedings of ICAD 2015, 9th International Conference on Axiomatic Design, edited by M. K. Thompson, D. Matt, A. Giorgetti, N. P. Suh, and P. Citti P., 137-142. Red Hook: Curran Associates.

Arcidiacono, G., Wang, J., Yang, K. (2015). Operating room adjusted utilization study. International Journal of Lean Six Sigma, Vol. 6, Issue 2; pp. 111–137

Bi, Z. M. (2011). Revisit System Architecture for Sustainable Manufacturing. Journal of Management Analytics; 3(9): 1323-1340.

Bi, Z. M., Cochran, D. S. (2014). Big Data Analytics with Applications. Journal of Management Analytics; 1(4): 249-265.

Bi, Z. M., Xu, L. D., Wang, C. (2014). Internet of Things for Enterprise Systems of Modern Manufacturing. IEEE Transactions on Industrial Information; 10(2): 1537–1546.

Dumbill, E. (2012). What is Big Data? An Introduction to the Big Data Landscape. O'Reilly Strata. http://strata.oreilly.com/2012/01/whatis- big-data.html.

Fogarty, D. (2015). Lean Six Sigma and Big Data: Continuing to Innovate and Optimize Business Processes. Journal of Management and Innovation; 1(2): 2-20.

Fogarty, D. (2015). Lean Six Sigma and Data Analytics: Integrating Complementary Activities. Global Journal of Advanced Research; Vol. 2, Issue 2

Franks, B. (2012). Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics. New York: Wiley.

Gobble, M. A. (2013). Big Data: The Next Big Thing in Innovation. Research and Technology Management; 56(1): 64-66.

Kaplan, F. (2015). A map for big data research in digital humanities. Front. Digit. Humanit. 2:1. doi: 10.3389/fdigh.2015.00001

Mo, Z., Li, Y. (2015). Research of Big Data Based on the Views of Technology and Application. American Journal of Industrial and business Management; 5: 192-197.

Parker, R. (2014). Big Data and Analytics in Manufacturing: Driving New Levels of Management productivity. IDC Manufacturing Insights; MI250786. Available from: http://www.tcs.com/SiteCollectionDocuments/White%20Papers/Big-Data-Analytics-Manufacturing-0914-1.pdf

Wang, L., Alexander, C. A. (2015). Big Data in Design and Manufacturing Engineering. American Journal of Engineering and Applied Sciences; 8(2): 223–232.

Risam, R. (2015). Beyond the Margins: Intersectionality and the Digital Humanities. http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html#transformdh2012

Svensson, P. (2009). Humanities computing as digital huminites. Digital Humanities Quaterly 3: 3. http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html

Unsworth, J. (2002). What is humanities computing and what is it not? In Jahrbuch für Computer philologie, Vol.4, Edited by G. Braungart, K. Eibl, and F. Jannidis, 71–84. Paderborn: Menis Verlag