Data and standardisation

I considered the potential applications of machine learning technology in the construction industry in a previous article 'Future of forecast: machine learning' when I reflected on how reliable forecasts and insights generated via machine learning technology may noticeably affect the efficiency of construction projects.

One of the challenges of using this emerging technology is the lack of efficient and reliable data[1] and therefore, the advancement of this technology relies on the determination of construction practitioners to record and standardise project data diligently, to support this core element of machine learning technology.

This article focuses on the necessity of capturing data into structured datasets - which are at the heart of machine learning - and explains the different data types, the importance of structured and standardised data and overall data quality.

Data types

The construction industry has a broad range of data types and sources.

Data, or information, can be anything; for instance, correspondences, photos, site diaries, reports, schedules and transactions. However, the data is stored on different platforms, such as emails, ERP Systems, spreadsheets, planning software, databases, etc. Unfortunately, in the majority of projects, data and data sources are not integrated with each other, which causes practitioners to duplicate work and waste time, but most importantly, to record the information in different sources under different formats. Recent research indicates that 96% of data generated in projects is not used at all and, data sources of 30% of engineering and construction companies are not integrated, with employees spending 13% of their working hours searching for information in data sources[2].

In literature, types of data are defined under the two main groups; structured; and unstructured data. Unstructured data is stored in undefined and native formats, and consists of videos, images, audio, emails, word and pdf documents, and project scheduling native files such as Ms Project (.mpp) and Primavera(.xer).

Currently, 90% of generated data (in construction and other industries) is classified as unstructured data, requiring data science expertise to get reliable predictions using machine learning algorithms[3].

[1] Curry, E. et al. (2021) 'Technical Research Priorities for Big Data', The Elements of Big Data Value. Springer International Publishing. doi:10.1007/978-3-030-68176-0_5

[2] Snyder, J., Menard, A. and Spare, N., n.d. Big Data = Big Questions for the Engineering and Construction Industry. FMI.

[3] Curry, E. et al. (2021) 'Technical Research Priorities for Big Data', The Elements of Big Data Value. Springer International Publishing. doi:10.1007/978-3-030-68176-0_5

Conversely, structured data is in a standardised format, allowing it to be quickly processed in databases or even MS Excel spreadsheets that are meaningfully organised, structured and appropriately checked and cleansed. Consequently, structured data can be used by the average business person and easily managed by a language, i.e. SQL (Structured Query Language)[4].

> "Data standardisation would allow companies to not only accelerate their digitisation and machine learning strategies, but also increase effective information communication with the potential to save 7.5% of the project's total expenditure"

Structured and standardised data

Considering one of the challenges in the adoption of machine learning applications in the construction industry is the availability of highly skilled people with strong domain, data analytics and data science knowledge, and of course, know-how[5], the importance of having structured and high-quality data is an imperative for the short and medium-term[6] as it is easier to manage and use by construction industry practitioners.

This then raises two questions: (1) how can the percentage of structured data in construction projects be increased?; and (2) how can the quality of data be increased and then maintained at a high level?

(1): One of the ways to increase the proportion of data that is captured into structured data is by converting it from unstructured to structured version, once it is first deemed to be valuable. In literature, this process is called ETL, which stands for Extract, Transform and Load. In this traditional process, firstly, the raw data is extracted from the data source, then the extracted data is reformatted and cleansed, and in the final stage, the reformatted data is loaded into the final target database[7]. As a practical example, the plain text is extracted from emails, then reformatted to .xml formats, and finally loaded to the final database.

However, this process requires data science expertise. This reinforces the need for companies across the construction industry to accelerate the training of practitioners and attracting skilled people capable of taking on such roles, so that data can be captured and turned into knowledge for subsequent re-use and application.

(2): The high quality of data impacts the accuracy of forecasts[8]; clearly therefore, it follows that low quality data would generate significant decision-making mistakes[9]. Therefore, ensuring that data is of a high quality is vital. Several researchers suggested different dimensions to assess and maintain data quality, the common ones being accuracy, validity, and completeness[10]. Similarly, Data Management Association

[4] Praveen, S. and Chandra, U., 2017. Influence of Structured, Semi- Structured, Unstructured data on various data models. International Journal of Scientific & Engineering Research, 8(12).

[5] Zillner, S. et al. (2021) 'A Roadmap to Drive Adoption of Data Ecosystems', The Elements of Big Data Value. Springer International Publishing. doi:10.1007/978-3-030-68176-0_3.

[6] McCord, S.E. et al. (2022) 'Ten practical questions to improve data quality', Rangelands. Elsevier BV. doi:10.1016/j.rala.2021.07.006.

[7] Saradava, H., Patel, A. and Aluvalu, R., 2016. A survey on ETL strategy for Unstructured Data in Data Warehouse using Big Data Analytics. In: International Conference on Research & Entrepreneurship.

[8] Sadiq, S. and Indulska, M. (2017) 'Open data: Quality over quantity', International Journal of Information Management. Elsevier BV. doi:10.1016/j.ijinfomgt.2017.01.003.

[9] Cai, L. and Zhu, Y. (2015) 'The Challenges of Data Quality and Data Quality Assessment in the Big Data Era', Data Science Journal. Ubiquity Press, Ltd. doi:10.5334/dsj-2015-002.

[10] Ramasamy, A. and Chowdhury, S. (2020) 'Big Data Quality Dimensions: A Systematic Literature Review', Journal of Information Systems and Technology Management. TECSI. doi:10.4301/s1807-1775202017003.

(DAMA) defines data quality dimensions under six main headings: completeness; uniqueness; consistency; timeliness; validity and accuracy[11].

There is no doubt that construction practitioners have the capacity to establish and maintain high-quality datasets, and one of the ways to increase the overall quality of data is “standardisation”. Recent research has identified that standardised records increase the quality of data[12] and therefore, the more standardised data input onto platforms, the more structured and high-quality data would be on hand for use by parties.

There are different ways to standardise information. For instance, during project schedule preparation, the data quality, project communication and productivity would increase if specific English vocabulary, classified words and adjunct words are used[13]. Moreover, the same logic of using specific words can be applied to site diaries, progress reports, quality and health and safety reports or even email topics. Additionally, tabulating, formatting, and integrating the data in the reports with one another would increase overall data quality and reduce human error[14]. Therefore, data standardisation would allow companies to not only accelerate their digitisation and machine learning strategies, but also increase effective information communication with the potential to save 7.5% of the project's total expenditure[15].

Summary

The amount of data generated daily is increasing exponentially with much of this data being unstructured, and therefore not easy to extract, analyse and use to provide valuable insight into construction activity. Considering the lack of highly skilled data analytic knowledge, it is not surprising that companies waste over 90% of data in the construction industry. It seems clear that construction practitioners would make great gains by standardising data inputs, integrating data sources, and working innovatively to prevent waste by reducing the volume of unusable data. This would eliminate errors, improve data quality and add value to our decision-making process and our industry.

In subsequent articles I will consider data ecosystems, common data environments, and whether data can be an economic asset.

Contact details

Oğuzhan Çinar
Senior Consultant
oguzhancinar@hka.com
T: +44 (0)7801 963 700

This publication presents the views, thoughts or opinions of the author and not necessarily those of HKA. Whilst we take every care to ensure the accuracy of this information at the time of publication, the content is not intended to deal with all aspects of the subject referred to, should not be relied upon and does not constitute advice of any kind. This publication is protected by copyright © 2022 HKA Global Ltd.

[11] DAMA UK, 2013. The Six Primary Dimensions For Data Quality Assessment.
[12] Ni, K. et al. (2019) ‘Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study’, BMJ Open. BMJ. doi:10.1136/bmjopen-2019-029314.
[13] [13] Li, C.F. (2012) ‘The Researches on the Standardization of Petroleum Exploration and Development Structured Data’, Advanced Materials Research. Trans Tech Publications, Ltd. doi:10.4028/www.scientific.net/amr.461.749.
[14] Curry, E. et al. (2021) ‘Technical Research Priorities for Big Data’, The Elements of Big Data Value. Springer International Publishing. doi:10.1007/978-3-030-68176-0_5.
[15] Hong, Y. et al. (2022) ‘Improving the accuracy of schedule information communication between humans and data’, Advanced Engineering Informatics. Elsevier BV. doi:10.1016/j.aei.2022.101645.