Living Lab

Transport Infrastructure Efficiency Strategy

TIES LIVING LAB PROGRAMME

Standardisation and classification of project cost data (IP5b)

October 2022

INTRODUCTION

Benchmarking is a fundamental tool for improving performance in the delivery of transport infrastructure projects. The sector has seen a significant boost in funding in recent years, and there is an urgent need to use benchmarking to inform decision-making processes so that projects benefit from collaboration and deliver future-proof infrastructure while ensuring value for money.

But benchmarking cannot take place without suitable standardised data – particularly data for assessing all aspects of project costs. Unfortunately, as the TIES Living Lab data research team (the Analytical Consortium) acknowledged from the outset, there is a lack of consistency in the way costs are reported across the construction supply chain and client organisations, making it very difficult to implement robust comparison within organisations, or more widely across the sector and internationally.

This information paper describes a project to demonstrate the possibilities of using artificial intelligence (AI) to extract, transform and classify project cost data in a standardised way using "data mining". The objective of the work, carried out under the project on Artificial Intelligence for Data Mining (and feeding in to the project on Metrics, Benchmarking & Repository) was to prove the concept in a "live" situation, taking data from a variety of sources and showing how AI can classify infrastructure project cost data into a common standard.

BACKGROUND

Infrastructure construction projects are notorious for the difficulties of controlling costs. This is not surprising, given the vastly different cost reporting systems and practices between organisations, between different parts of the same organisation, and between organisations and their supply chains. For example, with projects ranging in size from a "simple" footbridge to an entire underground station, there are numerous ways of preparing cost breakdown structures and various data collection techniques, so there are likely to be discrepancies in the granularity and classification of cost data. Individual organisations may have workflows and classification systems for agglomerating cost data, but these tend to be bespoke to the organisation and involve considerable manual handling of sometimes large amounts of data – requiring a very significant time commitment and professional or

personal judgement – and often resulting in cumbersome, error-prone and inconsistent data handling processes.

Nevertheless, cost data is always available to someone at the most granular level because it is money paid out for labour, materials, professional services and so on. However, fundamental cost data is often not available to the client, which is only provided with data at the contract level. To obtain the detail, clients need to specify their requirements in contract documentation.

Key industry players globally and, in particular, professional institutions such as the Royal Institution of Chartered Surveyors (RICS), advocate using the International Cost Measurement Standard (ICMS), which can deliver benefits such as better data portability and improved cost reporting, and hence better understanding of project outcomes and easier performance benchmarking (see ICMS 3rd edition, November 2021, published by the ICMS Coalition).

CHALLENGES

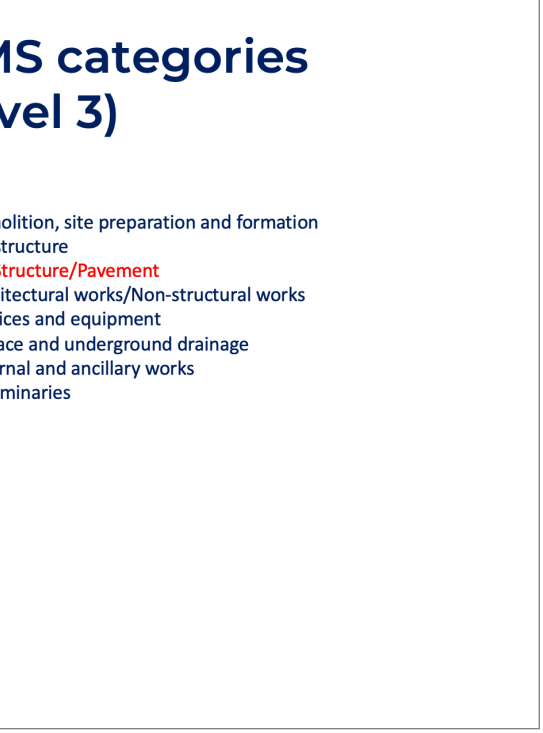
The ICMS provides a framework for classifying project cost data that will allow effective comparison across sectors and internationally. The ICMS framework can be interrogated at a high level (e.g. a bridge or a station) of data but, fundamentally, the ICMS structure is an elemental format. For example, data documents produced in elemental formats would normally include sections or headings such as “preliminaries”, “substructure”, “superstructure” or “services and equipment”.

There are several challenges when gathering data and implementing the ICMS framework:

- The ICMS is more effective for classifying data in an elemental structure, but several transport infrastructure organisations use other methods that are more closely linked with tracking expenditure, such as trades-based or asset-based classifications. This influences the headings and subheadings in cost documents and therefore poses an additional challenge for consistent data extraction.
- There are also variations in the granularity of the available data. Some costing documents record only high-level costs, and it can be difficult (if not impossible) to map these to other cost breakdown structures. For example, documents with aggregated costs for, say, “concrete works” cannot be broken down further to elements such as substructure and superstructure, if the underlying granular data is not available.
- Transport sector organisations have a fair amount of historical cost data in standard formats that can be aligned with the ICMS methodology. However, the data needs to be re-classified and organised into a format that is more compatible with the principles of the ICMS. Manual reclassification can be subjective and error-prone.
- There are no standardised tabular structures (e.g. headings in spreadsheets and databases) used for recording cost data across the transport infrastructure sector or even for different parts within the same organisation, which poses a challenge for automatic data extraction using AI techniques.

DEMONSTRATING THE POWER OF AI

As part of the TIES Living Lab Programme, the project on Artificial Intelligence for Data Mining focused on developing and testing an AI-based tool for analysing historical cost data. The AI tool – known as the ICMS Cost Classifier – learns from manually classified data and then attempts to classify any new entries provided.



For this project, data was provided by National Highways (NH), covering 35 major projects relating to changes in motorways (e.g. conversion of a standard motorway into a smart motorway), and other projects such as renewals costs breakdowns for maintenance. The dataset consisted of over 50,000 cost descriptions. One important advantage of using NH data was that the internal NH "Work Breakdown Structure" (WBS) is compatible with ICMS, and NH had already developed a map between its WBS and the ICMS framework.

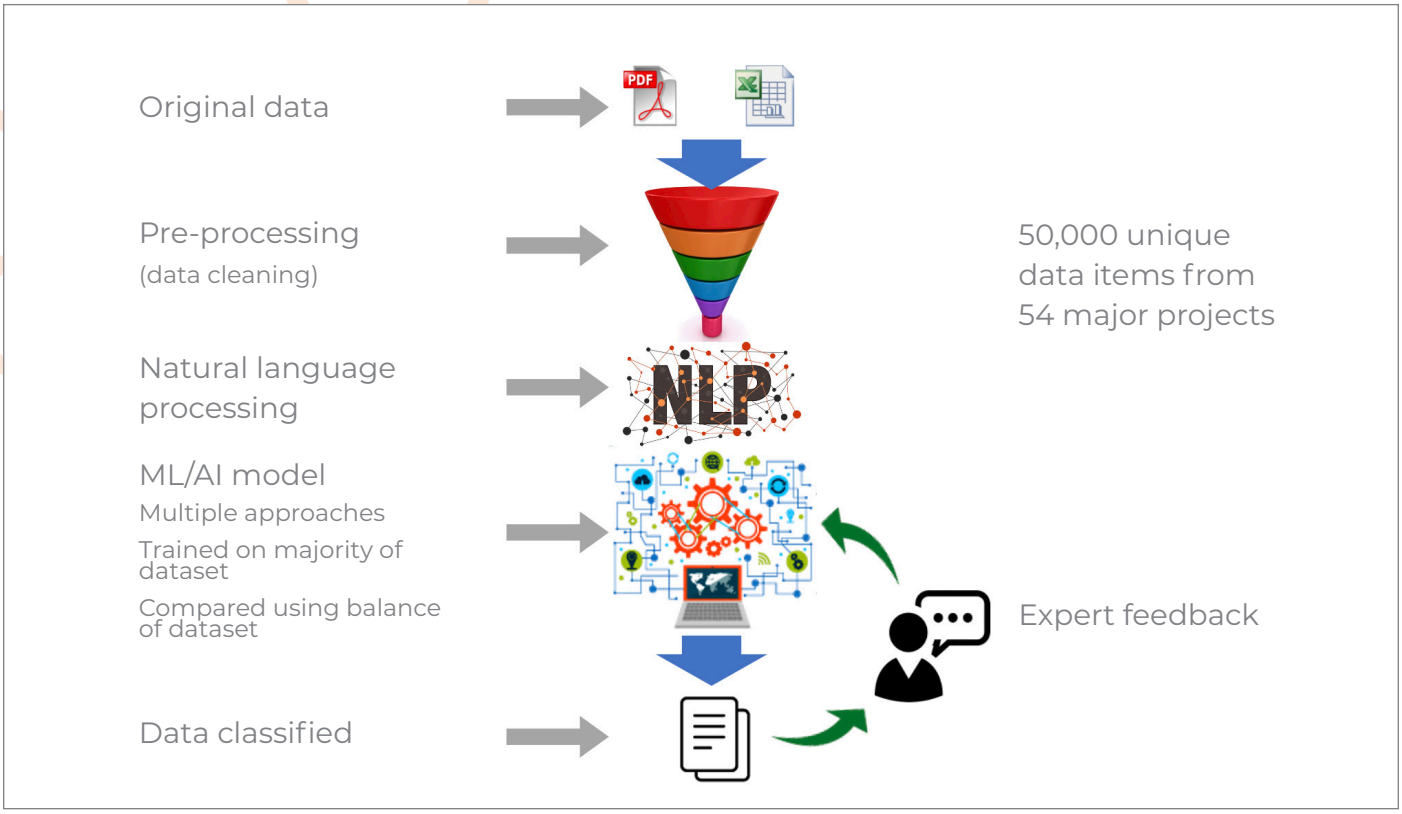
Classifying large quantities of historical cost data

The novel AI-based tool is based on a combination of natural language processing (NLP) and a variety of machine learning (ML) algorithms, and can automatically classify cost descriptions into their appropriate ICMS categories, which can subsequently be manually verified as required.

The AI approach starts with automated data extraction and transformation (the semi-automated ICMS Cost Classifier).

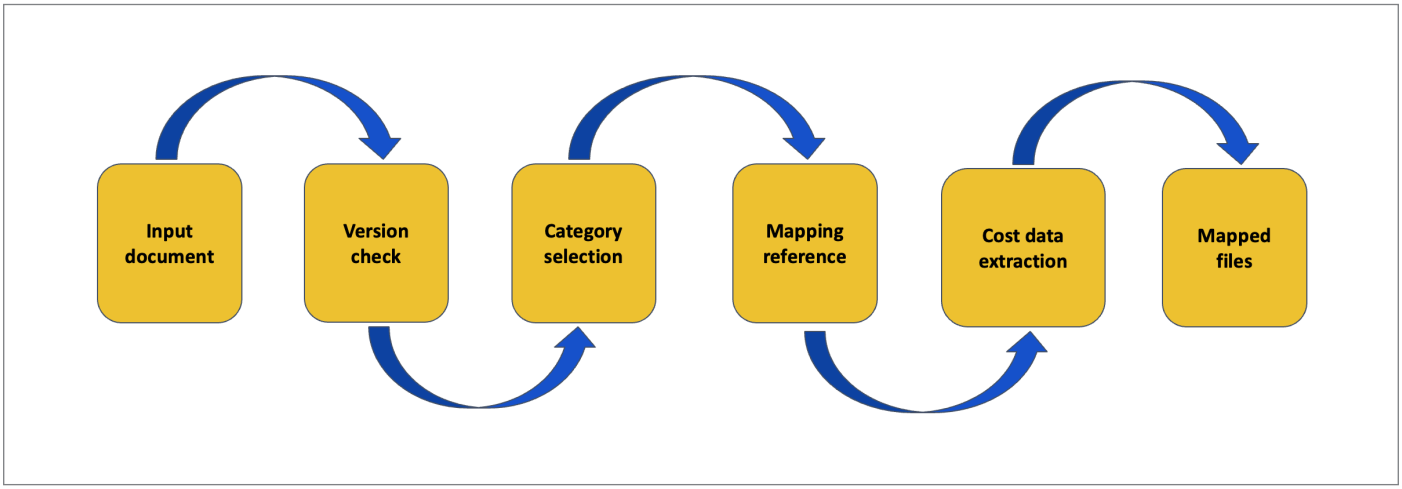
This transforms "text" into "vectors" (i.e. pieces of data with associated attributes/values such as price in £, or quantity in m^3) and then uses ML to find the optimal vectorisation method (e.g. binary, count, frequency), including "term frequency to inverse document frequency" (TF-IDF) approaches. This is then combined with the optimal ML/AI techniques to obtain the desired classification, namely: naive Bayes (as a baseline predictor), and then Random Forests, Support Vector Classifier, and finally multi-layer perceptron (deep learning) with a variety of optimisations until the best accuracy was obtained.

An overview of the process is presented below.



Handling inconsistencies

A specific challenge is that cost data is often not recorded consistently across the transport infrastructure sector. For instance, within NH, the Major Projects Team uses a different document structure for cost reporting to that used by the Renewals Projects Team. Potentially, this poses a further challenge for automatic data extraction using AI techniques, although such differences can be easily overcome providing there are not large numbers of reporting templates. A solution tested in this project was to develop a different data extraction pipeline for each cost reporting document template (illustrated below).



A step-change in accuracy and processing time

When tested on the NH data the AI tool achieved 80–95% accuracy, and cut the processing time by up to 90% (depending on volume of work) compared with the time needed to do the same work manually – a step-change in terms of accuracy and processing time compared with traditional analysis methods.

The tool allows input of cost descriptions either as a single item or as a batch of several descriptions in a spreadsheet and returns an output of the description with a suggested ICMS classification.

This tool has been deployed on a cloud-based platform (see https://icms-classifier.co.uk/), and has the potential to assist data and cost intelligence teams in benchmarking and data mining in the following scenarios:

- Classifying historical data to obtain backwards compatibility inside and across companies
- Checking already classified cost descriptions for accuracy, before further analysis.

LESSONS LEARNED

- Data should be captured at an elemental level, but, provided the cost data can be captured at a granular level, translating the data into an ICMS format is generally feasible.
- Data captured and stored should be at a granular and elemental level, to improve opportunities for cross-mapping while maintaining data relevance for internal purposes.
- Using industry-wide schemas and templates for data storage (e.g. in spreadsheets) would pave the way for AI-driven data extraction, transformation and loading to enable speedy and automated data mining.
- Cost descriptors sometimes contain typographical errors, and various abbreviations are used by different cost professionals and teams or even in the supply chain to represent the same item. There is scope for future development of the AI tools to deal with this.
- AI performs better with larger quantities of data. In many instances data availability was limited, hampering application of AI techniques. For example, some of the project data received from arm's length bodies as part of the TIES Living Lab Programme were in the tens and hundreds.
- There is an imbalance of available data. While some cost items are popular and recurring others are seldom used, making AI learning in the under-utilised categories more challenging.
- Some aspects of cost data classification into a common standard remain subjective, and even experts may have different opinions. For example, the input data "Trial holes on first night of works to prove service locations: allow excavation gang and all resources to undertake bituminous carriageway reinstatement" was classified by two experts to the ICMS codes 1.02.020 and 1.01.010. This is tricky because the WBS code indicates this as part of "earthworks", while another reviewer thought the item a better fit with ICMS code 1.01.010 on "site survey and investigation".

- This problem has been acknowledged by providing a feedback mechanism that allows users of the automated process to question and challenge output as well as provide feedback that may be incorporated in future iterations of the tool.

LEGACY

This preliminary work has demonstrated very promising results on the applicability and accuracy of AI-driven data mining, while also highlighting ongoing challenges to the wider use of AI for standardising cost data for benchmarking.

This "proof of concept" approach also paves the way for automated classification of other project data types that have been written in a natural language, such as risk, environmental data or site diary entries and reports, with the opportunity for significant savings in terms of time, expertise involvement and unbiased classification

The ICMS Cost Classifier performs very well on NH-related data inputs, but it will require additional training for other data types (e.g. for the rail sector) to improve cross-organisational usage. Additional models may be developed and trained for each sector, when the opportunity arises beyond the TIES Living Lab Programme.

This work was led by Professor Lamine Mahdjoubi (Professor of Digital Built Environment) of University of the West of England (UWE Bristol), and overseen by the Analytical Consortium under the project on Artificial Intelligence for Data Mining.

Living Lab

Transport Infrastructure Efficiency Strategy

The TIES Living Lab is a transformative collaboration of 25 partners together with Government, i3P and the Construction Innovation Hub that use data, technology and Modern Methods of Construction within live transport infrastructure projects to deliver significant value-adding benefits across the transport infrastructure sector. The programme is funded via a grant from Innovate UK through the Transforming Construction programme, plus contributions from the Department for Transport, HS2, Transport for London, Network Rail and National Highways.

The four strategic outcomes of the collaboration are to:

1. Improve the way Transport Infrastructure projects are set up to maximise value
2. Achieve better assurance of project and programme value and what assets should cost (benchmarking)
3. Accelerate the wider adoption of MMC
4. Establish the TIES Living Lab as a catalyst for long term cultural change across sectors by making a compelling case for long term HM Treasury funding to scale this facility.

Project led by: The National Skills Academy RAIL

Project sponsored by: Department for Transport; Industrial Strategy; UK Research and Innovation

Bryden Wood; Akerlof; Kier; Accelar; WLC Whole Life Consultants Ltd; Network Rail; SRC Infrastructure; loop The Social Value People; East West Rail; Office for National Statistics; X-treme Systems; national highways; HS2; Transport for London; University of Leeds; mtc Manufacturing Technology Centre; Walker Construction; RICS; mavin POWERCUBE; UWE Bristol University of the West of England; University of Dundee; expedition; COSTAIN

Published by RICS on behalf of TIES Living Lab (IP5b)
Project led by NSAR, with programme management support from Limberger Associates