Using AI to Bring Dark Data to Light

Submitted by Stephen Bourne, P.E., Project Director, SNC-Lavalin/Atkins

"Dark Data – The information assets organizations collect, process, and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships, and direct monetizing)." Gartner IT Glossary

If a tree falls in the forest and there's no-one around to hear it, does it make a sound? If an as-built drawing collects dust in a records room and there's no-one to digitize it, did it ever exist? As paper, vellum, and microfiche records continue to decay, organizations struggle to preserver this dark data and extract its value. Dark data is non-digitized data; artifacts of the pre-digital age whose existence is tied to the robustness of the media on which they are recorded. One fire. One spilled cup of coffee. One accidental trash dump, and dark data is gone.

The solution is to convert to dark data to digital form. In the process, we can integrate the data with other digital data on the same assets, to produce a rich digital record of the assets described.

The problem with dark data is that there is too much of it. Records rooms with thousands of poorly organized drawings are a commonality. Converting all that to digital form takes way too much effort. Integrating within your enterprise databases takes even more effort. That's where artificial intelligence (AI) can help.

The R&D team of SNC-Lavalin's Atkins business decided to tackle this problem by creating a new AI-based tool called InfoExtract.

Think Like a Human

InfoExtract is built to match the behavior of a human being. As extractors, our human toolbox includes the abilities to:

- **Read and Comprehend text** – words, phrases and sentences, paragraphs, and sets of paragraphs that collectively make a point.
- **Understand visual depictions of real-world structures** – drawings that depict multiple assets from multiple systems in a single paper space, or may depict only a detail of a larger asset. It also means understanding how to link the drawing in paper space to a real-world location
- **Anticipate how documents are organized** – knowing the intent of the document is half the battle. If we know the template a set of documents follows beforehand, we are empowered with knowledge of content to expect within the document.
- **Find the signal in the noise** – humans can see pattern where machines often fail. "Captcha" validation technology has shown this clearly; we can see the letters "JFXN9", where machines stumble on deliberate obfuscation. An analogous skill is seeing the structural components in a drawing, even though they share paper space with electrical, plumbing, labels, and so on.
- **An uncanny ability to employ these skills together to understand the document** – our most valuable skill is knowing the combination of and sequence with which we should apply our skills to extract what a document is saying.

With InfoExtract, Atkins has attempted to create the same set of abilities within the software, and the overriding "meta-ability" to choose when to use them, empowering the tool to read any set of documents it is presented.

Case Study

Recently, Atkins did some work parsing drawings of rail stations in a major city, which had 18 months to extract data from upwards of 500,000 drawings of stations and rail sections. As a proof of concept, InfoExtract was given the objective of searching for escalators within the drawings and adding those it finds to an Esri geodatabase of assets. The geodatabase started as what was called an Info-Scaffold, a simple stations map layer with a single polygon for each station and the station name as the single attribute.

InfoExtract decided to use the following algorithm as it parsed the drawings. Note that the only human input was the name of folder that held the drawings and the fact that we were looking for escalators.

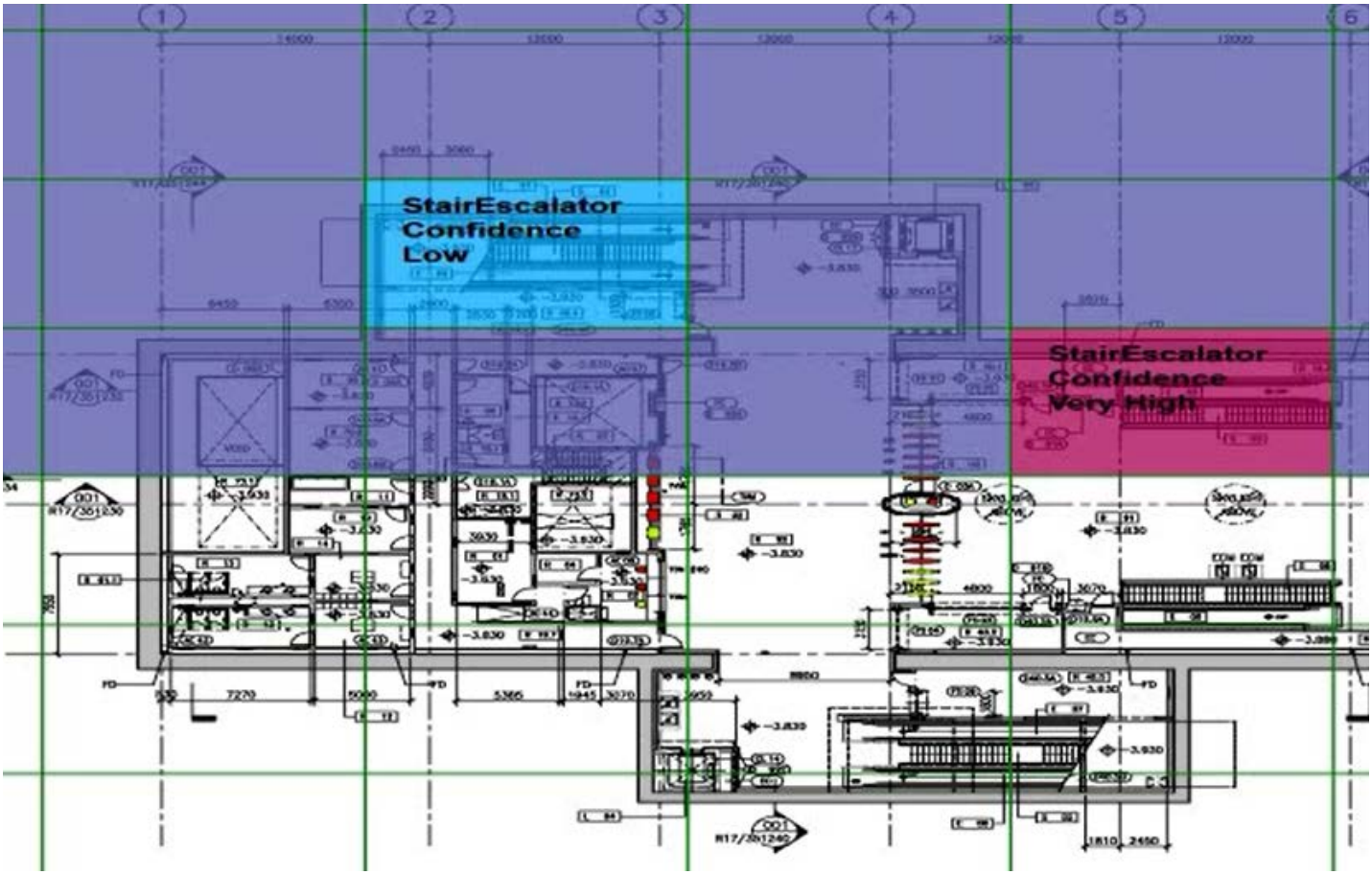

Figure 1: InfoExtract searches a rail station for stairs and escalators.

1. Search the drawing for a title block and use optical character recognition (OCR) to try to find a set of words that is the same as the name of a station in the Info-Scaffold. If any assets are found, they will be associated with the station that is identified.
2. Use a suite of “shake the trees” algorithms which overcome poor OCR performance, often due to horizontal and vertical lines that form the borders of the title block, or stains or fading text present in older drawings.
3. Slice the drawing up into tiles and convert the tiles into collections of vector lines. Compare the statistics of the vector line collections to statistical patterns of known assets. Tag those tiles with patterns similar to the escalator pattern.
4. Try to improve the signal to noise ratio by iteratively taking images within the region of the tile and repeating the pattern comparison.
5. If above a confidence threshold, store the asset in the Info-Scaffold and associate it with the station found in item 1.

Note that this algorithm was chosen based on the documents InfoExtract was being provided, and the objectives we set for it. Had the documents been text reports, InfoExtract would have selected a different algorithm more suited to the information present in standard technical reports.

In this case study, InfoExtract was 85% accurate in finding all the escalators and elevators in the set of drawings we used.

AI and Experts Working Together

AI shouldn't work in a vacuum. The case study scored 85% accuracy with little supervision from the human operator. A better approach would be to have the tool present low confidence results to the operator and have the tool learn as the expert provides the correct answer. For example, we saw in one case that InfoExtract thought the parallel lines in table embedder in a drawing were an escalator. Given this example to correct, the expert would have instructed InfoExtract to first search for tables and remove areas with them from the search.

The Future

SNC-Lavalin/Atkins' R&D group continues to leverage AI concepts to bring value to the asset management space. We will enhance InfoExtract to produce rich asset databases of stormwater, rail, energy, utilities, and other asset classes.