Dstl releases free Baleen 3 data processing update

The Defence Science and Technology Laboratory (Dstl) has released a new free version of its popular data processing tool.

Baleen 3 is a tool for building data processing pipelines using the open source Annot8 framework and succeeds Baleen 2, one of the first open source projects by Dstl, the science inside UK defence and security. It offers users the ability to search, process and collate data, and is suitable for personal and commercial applications. It has been used across government, and by industry and academia, and also internationally as well as in the UK.

The tool enables the creation of a bespoke chain of “processors” to extract information from unstructured data (e.g. text documents, images). For example, Baleen 3 could process a folder with thousands of Word Documents and PDFs in it to extract all e-mail addresses and phone numbers in those documents and store them in a database.

As well as text, Baleen 3 can also find and extract images within those documents, perform OCR to find text within those images, translate that text into English, and then run machine learning models to find mentions of People within those images.

Baleen 3 supports components developed within the Annot8 framework, and as a result it is easy to extend and develop further to cover new use cases and provide additional functionality. There are already a large number of components available for use within the Annot8 framework, including some previously developed by Dstl.

Following the release of Baleen 3, support for the existing Baleen 2 project will be withdrawn. Dstl is encouraging all users to move to using Baleen 3 where possible. Baleen 3 is built on top of newer technologies, and will be easier to maintain and deploy as a result of the upgrade. It also extends Baleen 2’s focus on text to support other forms of unstructured data, such as images.

Baleen 3 was co-developed by Dstl staff on the Augmenting the Analyst project (Information Systems programme), and Committed Software.

Baleen 3 is available to download now from https://github.com/dstl/baleen3. For more information, please contact oss@dstl.gov.uk.