Market Need

Many businesses store pools of information-rich text in their systems - such as reports, customer reviews, user comments and general word processing documents.

Insights can be uncovered by analysing these text sources to discover underlying hidden topics or trends, such as unexpected clusters of words across documents.

In reality, however, it can be difficult to manually examine all this of information to discover such hidden trends or topics.

Figure 1. Visualising document groups based on word clusters or topics

Technology solution

DocoPool - a web tool that allows users to explore the content of text documents for hidden knowledge:

- Identifies and visualises word groupings or "topics" across sets of text documents

- Each document is carved up into individual words and word frequencies

- Uses a probabilistic topic modelling algorithm to discover the spread of word occurrences across a corpus of text documents.
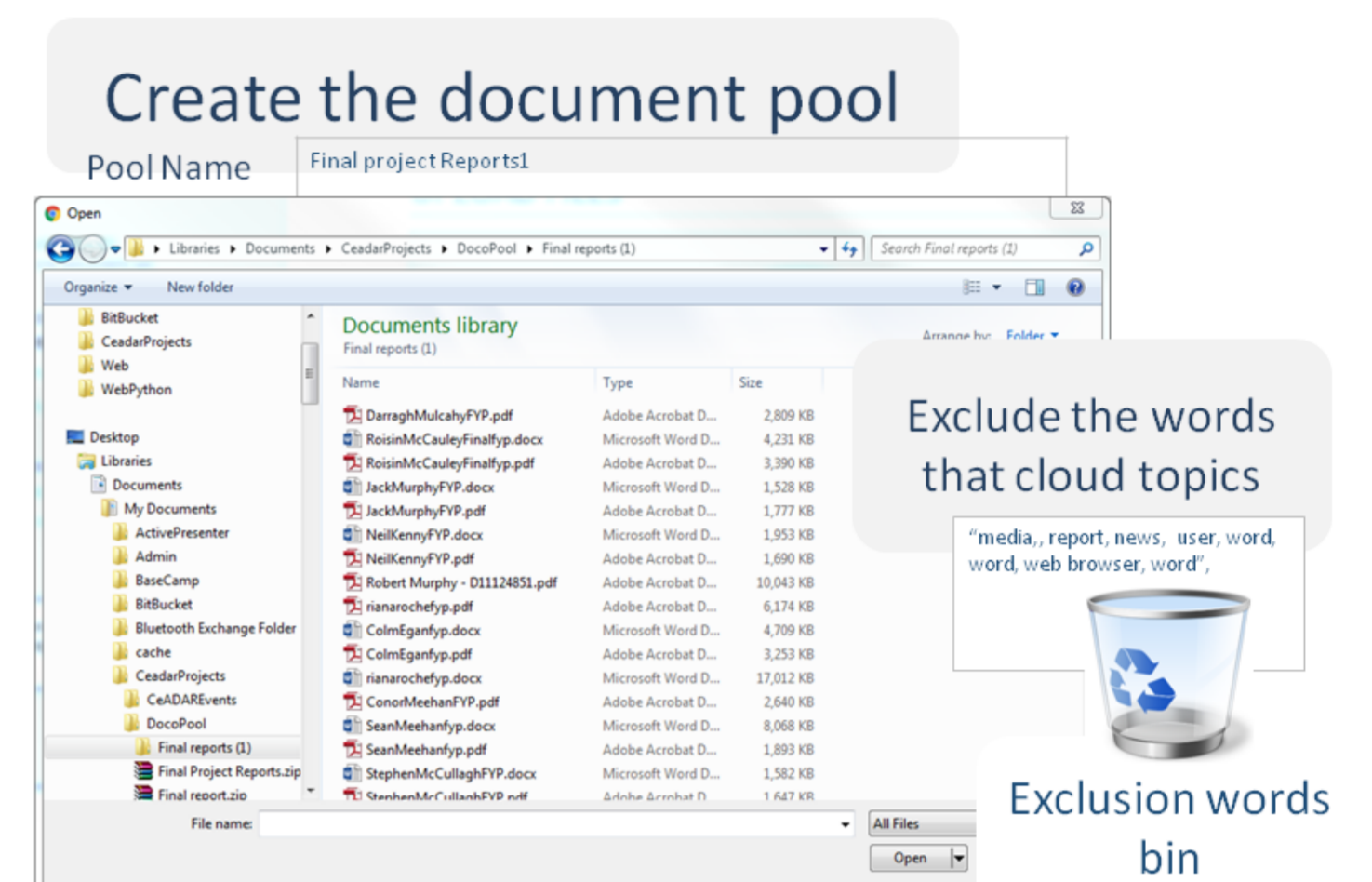

Figure 2: Users select their documents for analysis, and define domain specific word exclusions

Key features

- Easy-to-interpret visualisations

- Drill-down on document details for deeper analysis of word clusters

- Specialist or domain specific word exclusions – to prevent clouding of hidden topics

- Flexible document upload (.txt, .pdf and .docx)
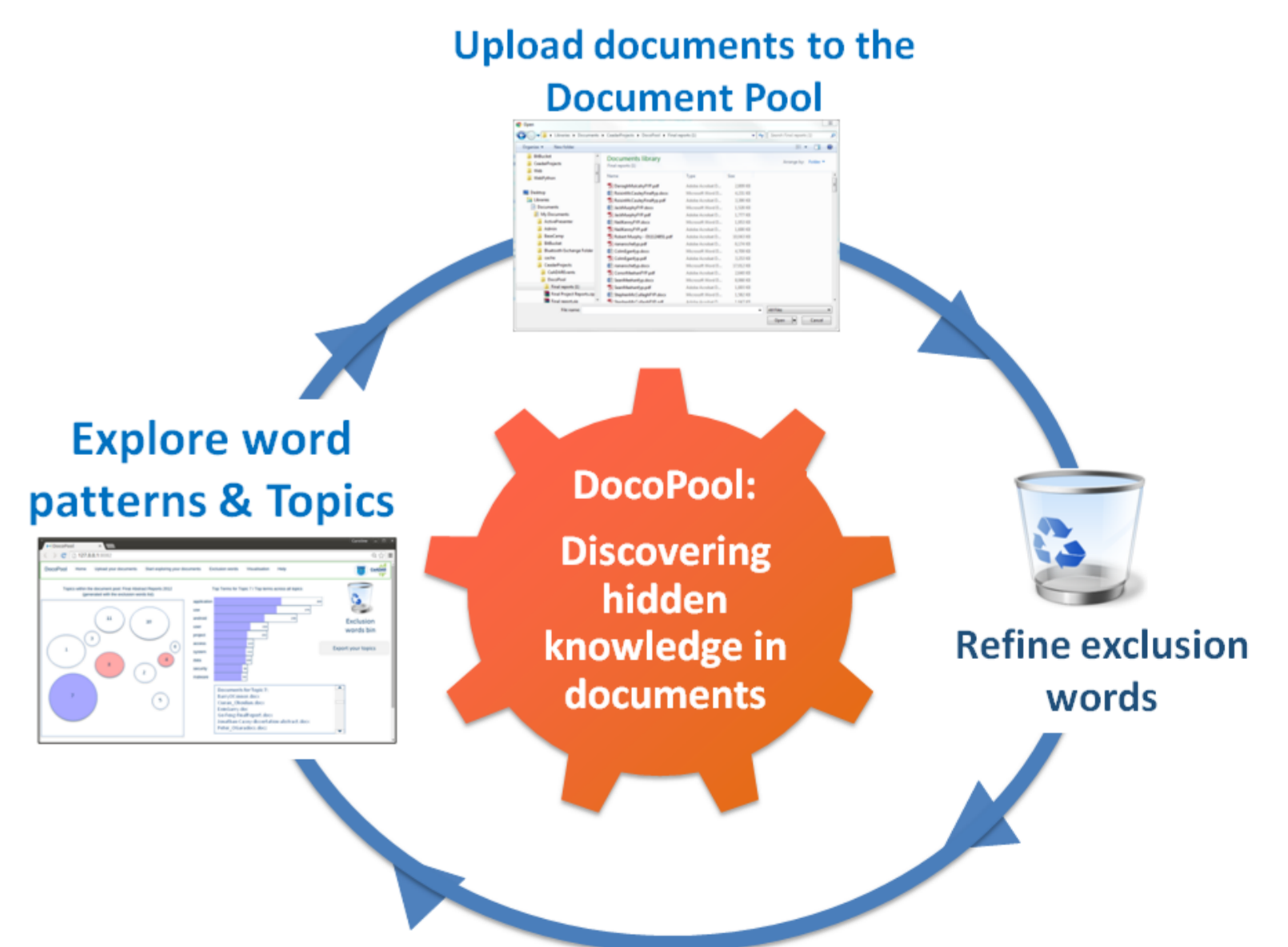
- 'Save' facilities to allow revisiting of explorations.

Figure 3: Document exploration: An iterative process

Research Team

Dr. Caroline Maillet Dublin Institute of Technology

Dr. Susan McKeever Dublin Institute of Technology

Follow **CeADAR Ireland**: