Requirements for 'PureC' persistent search platform

Julian Ellison, Tablane Technology, 6th July 2007.

Headline Requirements

The intention of this project is to create a server-based platform which can conduct persistent searches on the Internet related to the content of documents hosted on the environment.

The platform will complement the 'transactional' or 'on demand' search paradigm delivered by Google and others today.

Thus the platform should provide an automated method for monitoring content that could relate to the document. The document could contain free text, as well as content that is clipped from Web sites, RSS feeds and discussion groups.

This persistent search service should not only monitor the content, but make it available within a Web based interface for the owner of the document to review, and, if appropriate, modify the contents of the document thereby.

Document Meta-Data

Clearly the meta-data format of the document needs to be considered carefully.

Tags may be applied to the document by the owners of the document according to their own taxonomy, and additional tags

may also be applied automatically based, perhaps, on natural language analysis of the contents of the document.

The origin of the clipped content should also be preserved in the document.

The specification of meta-data relating to all, and parts of the document, is a requirement of the project. This meta-data will be required to enable the hosting platform to conduct its persistent searches, identifying servers to be monitored, keywords and tags to recognise, discussion threads to follow and report back on and indeed any other method that is viable.

An important area of investigation to be considered is how 3rd party content sources that are not already included in meta-data are discovered and searched for relevance. In other words, does the environment only search persistently for content from known sources, or can it augment these known sources with new sources it discovers.

Document Formatting

The specification of this meta-data will be used to extend a document management tool already created, known as TClipper (available from download.com, search for Tablane).

It is expected that while this tool will be the first tool of its kind to enable documents to be formatted in the correct manner for the environment, the specification will be opensourced to allow 3rd party developers to align their own tools to the specification.

It may be also that the environment itself can accept html content and apply the meta-data itself.

The document will not necessarily be a single, flat file. It will almost certainly be presented in a tree structure. The Collection concept within TClipper, with multiple 'Pages' demonstrates what is intended.

Document Ownership

The owner of the document should be able to determine whether the document is for private or public use.

It should also be possible to invite other registered members to access the document, and to contribute content to it. Thus the document owner should be able to authorise members of a group to have read or read/write permissions.

Modifications to the document should be notified to those able to view the document via RSS.

Interface Design

The interface design is clearly important. The intention here is to provide an uncluttered interface aligned to contemporary design trends, which ranks new sources of relevant content according to parameters we need to define, but which might include types of source (discussion groups, web sites, RSS feeds) and ideally relevance. It should be possible to request the environment to ignore sources in future, or to look more closely at new related sources of information.

It should be possible to view, within parallel windows, the document and the retrieved sources of information so that visual comparisons can be made. Methods of dragging new content into the document from the interface and editing the document within the interface would be useful.

Achieving the Prototype

The wish list of functionality here is obviously very broad. Given the resource and time constraints, urgent consideration needs to be given to what is achievable, and what makes best use of existing technology developed by DERI. It may be that the optimal approach is to concentrate on showcasing these technologies within this context, and separating out clear areas for further research and development once the project achieves greater funding.