Fusepool (SME) for Information Pooling for Product/Service Development and Research

Data is an integral element of the business value chain. However, many barriers impede leveraging the full potential of that data across existing data silos. Fusepool crosses borders to provide data for better product/ service development and research:

- PartnerMatch: How to find the right business partners?
- FundingFinder: Where to get funding for research?
- PatentExplorer: What is patented in my product area?
- JournalExplorer: What prior art exists in my product area?

These applications run on an open-source cloud platform for end users. The goal is to help small and medium enterprises (SMEs) better adapt to and exploit technological and regulatory changes. Fusepool digests large datasets from a variety of sources, such as patents, scientific articles, funding opportunities, standards and regulations. With its focus on SMEs, Fusepool is intuitively usable via graphic user interfaces that adjust to different needs. Currently, Fusepool is focusing on the medical and life sciences domains but can be extended based on the data that is brought into it.

Relevant concepts are uniquely identified in Fusepool. Almost all of these concepts were already defined elsewhere, for example using the Disease Ontology database of the University of Maryland School of Medicine, which in turn links its concepts to well-established and well-adopted terminologies such as the Systematised Nomenclature of Medicine (SNOMED), International Statistical Classification of Diseases and Related Health Problems (ICD), Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS). If Fusepool detects these concepts in patents or articles, they are instantly linked with the same concepts around the globe. That's the power of the semantic web: if we can agree and share minimal standards, software can start connecting us in more meaningful ways.

In the first flagship showcases, Fusepool digests several data sources, such as scientific articles and patents. We're very thankful to the team maintaining MEDLINE/PubMed at the US National Library of Medicine (NLM). They provide scientific articles on a daily basis in a common Extensible Markup Language (XML) format as well as ongoing support regarding suitable data sources. We're also thankful for the millions of patents from different sources that are normalised to a common XML format by the Information Retrieval Facility (IRF) hosted at the Information and Software Engineering Group (IFS) in the Vienna University of Technology. We're currently working with National Contact Points to make relevant data from the EU's research arm accessible in common XML format.

One example of is the Disease Ontology database for identifying diseases within biomedical data digested by Fusepool. Maintained by the University of Maryland's School of Medicine, the Disease Ontology can help in improving the aetiology and pathogenesis of diseases and identifying research on the best therapeutic options available. It provides and links to well-defined information, thereby giving context to

the data and increasing the value of the original data through crosslinking. To maximise these capabilities, the Disease Ontology could be combined with gene data using the Gene Ontology, for more comprehensive dataconnectivity and crosslinking.

Based on semantics and Linked Open Data (LOD; e.g. patents and publications), Fusepool is adding user feedback to derive optimisations for specific user groups. Enabling users to access metadata beyond the capabilities and recommended principles of the World Wide Web Consortium (W3C), Fusepool's method has amplified the possibility of global content sharing.

A multi-step search using keywords is inefficient, for example if you are only interested in documents to find authors, because search terms rarely provide access to highly specific, unknown or uncontextualised information. In contrast, the semantic web adds meaning to data, allowing the targeting of specific data by creating a context, relating the term to other relevant information in the field. Fusepool recognises the need to search for information within a specific context (information mining based on shared concepts, such as diseases or genes), which is extremely beneficial to medical researchers and practitioners, as their work requires comprehensive yet very accurate data.

The platform is now capable of providing autocomplete searching and faceted browsing, facilitating the search and retrieval of information across different data silos. Through these integrated apps based on Fusepool cross-referencing and data interlinking, not only can the user find more details on the search term, but classifiers can be created and reused to find the most relevant results for the individual or project. Moreover, Fusepool is implementing tools capable of creating contextualised landscape maps that supply visual analytics for large datasets.

Fusepool released the first major prototype of the software platform with most of the core functionalities of the final release, which include search and retrieval across all content, faceted browsing of structured data and user feedback collection, such as thumbs up/down. In parallel, some advanced functionalities are being developed, such as a network navigator for browsing a graph, landscape mapping using text similarity, data boosting for reweighing data sources based on preferences, label recommender for suggesting likely tags and adaptive graphical user interfaces that render semantically marked information according to user preferences.

Fusepool is an ecosystem that encompasses the data value chain. We still undervalue the role of data. For example, finding something that is not known is nearly impossible with keywords, and data are scattered around, unconnected and lifeless; a USB stick here, a CD-ROM there, some data on hard disk, others in a database and some in the cloud. Data, like information and knowledge, need to be connected to create value. When we look back in 40 or 50 years, how we treated data in the age of big data will look like how stones were treated in the StoneAge – we are just discovering the true potential of data.

The open-source data platform and business ecosystem has been developed in collaboration with end users, researchers, requirements engineers, data providers, data refiners and platform providers, in order to create a comprehensive ecosystem. The project is funded in part by the EU Seventh Framework Programme (FP7) under grant number 296192.