How-To Build A Data Catalog using Semantic Logging

In a data driven world we have to navigate along new paths. Big Data is the magic which opens up new dimensions. This can be a challenge besides a huge advantage. Not much was done in the past to map the "data world". Search technology was driving the usage of the internet but search alone gives us no strategic orientation in the data space. Equally, having paper and ink does not mean that you have a map for your journey already in your hands.

Like a traditional map needs paper and ink a mobile device needs appropriate apps in our modern world – but the content is what matters most, followed by usability aspects. This is why data catalogs should not be evaluated by completeness and accuracy only. Some means of fuzzy search and even fuzzy information is sometimes better than having no information at all.

Let's look deeper into usability. A data catalog should be open in terms of technology and access to it's content. Since the catalog contains data about the data - not the data itself - we shouldn't be too picky. But as always – being careful is better than too much of risk, especially in a very open world. With this in mind, we should say: "The data inside the data catalog should be as open as possible for authorized people and systems." We have to build up the data catalog without technical limitations which could stop us from exporting and merging multiple catalogs.

Time evolution of the content matters as well! Data sets are usually not static as books are – especially in near real time scenarios (driven by an IoT world) we have to consider the life cycle stages.

The Etosha project builds the open source components for open data catalogs. This means that data inspectors and profilers for integration in your own application or simply for interactive use in the Spark-shell build the base of the project. Full-text search, a graph database, and an RDF store build the backbone for multiple access strategies. RDF was chosen as data export and exchange format. Finally, the success of Git technology motivates us to offer a rich set of publication and integration capabilities.

The following section addresses some issues, raised by Todd Goldman in his blog [1]:

Discovery and Automated Tagging

Due to the large amount of information and data which surrounds us we need an automatic approach beside crowd-sourcing.
Etosha offers data inspection procedures and routines for automatic execution during data ingestion or directly afterwards. Footprints of individual columns of structured data sets are extracted for classification and comparison with future datasets.

Crowdsourcing

Especially for collaborative projects it is important to share knowledge and to learn from others. This means, integration of knowledge contributed by others is an essential functionality.
Etosha allows export and import of metadata and data models using the RDF data format.

Ratings and Reviews

Automatic data extraction can lead to non accurate or even completely wrong results. Users, especially humans have to evaluate and correct such results. Reviews, ratings provided by humans combined with automatic suggestions and cross-validation procedures provide the quality assurance we need in an open data catalog.

*Etosha uses an open feedback system – integrated into Github. In the future, we want to establish the **Global Data Map**: GDM is a specific view on a data catalog (or on multiple aggregated data catalogs) which provides context to data, such as dataset usage, dataset ratings, and user feedback.*

Integration Interfaces

Remember: content is what matters. The catalog provides information and knowledge for a variety of tools in multiple categories. From business glossaries to automatic ETL pipeline generation up to data governance and risk analysis – a lot of potential is already in your data catalog.

Etosha tools are focused on providing the best open source data catalog for the data driven business. It is on you to request the information. If it is not there, we have to find a way to provide the facts.

Etosha Data Catalog is More than just Search:

Since search-like interaction is a common feature, we decided to build Etosha around a multi-layer search interface. The search dashboard is the starting point for your work with the Etosha data catalog. We simply replaced the good old menu bar by some stored searches for quick reports, enriched by multiple search facets and visuals. Furthermore, semantic search using SPARQL and the and graph query language Cypher add more flexibility to your search experience. All of this has one goal: make your data catalog accessible.

[1] http://blog.waterlinedata.com/blog/search-is-not-a-data-catalog?utm_content=52869351&utm_medium=social&utm_source=linkedin)