Investigating Corruption with Graph Relationships

RTÉ, one of Ireland's most prominent broadcasting companies, has been at the forefront of Irish media for the better part of a century. Providing content through different mediums, they are no stranger, nor do they shy away from controversy. The RTÉ Investigations Unit is a branch established to provide journalistic investigative output on a multi-media basis for television, radio and online platforms for the Irish national broadcaster. When tasked to investigate public representatives and their purported assets, it needed to form links between seemingly unstructured and disconnected data to uncover undisclosed assets and expose their respective representatives. The result of this work culminated in an award-winning documentary, shedding light on the widespread under-declaration of assets by politicians.

Structuring Data

Upon performing extensive research into potential undisclosed assets of elected public officials, RTÉ Investigations Unit had to accurately and extensively test and query matches from the public property and corporate ownership registries against the interests declared by elected public representatives. This meant linking public records data with incomplete yet structured data gathered on the elected officials.

This process began with the creation of an ETL (Extract-Transform-Load) module, enabling semi-structured data from different sources to be cleaned up and ingested before being stored in a database. Once this heavy lifting process was completed, a database was needed to identify, visualize and uncover relationships between datasets.

Leveraging Multi-Model Capabilities to Exploit Graph Relationships

Upon having researched different vendors on the market, developers for the Investigations Unit found that most of these were equipped with decent document features, though OrientDB excelled at handling relationships between entities.

Industry
Media / Journalism

Challenge
Analyze and link both structured and unstructured data to uncover links between declared and undeclared assets by politicians.

Approach
Connect documents using an OrientDB graph to then be traversed. Exploit edges to emphasize links between datasets. Use OrientDB Studio for exploratory analysis on the graphs.

Solution
Utilize OrientDB's driver-rich catalogue and SQL support. Leverage multi-model capabilities by converting data matches to edges in order to further explore possible relationships between data.

Result
Clear visible links between seemingly disconnected data, uncovering widespread undisclosed assets by politicians and exposing elected representatives apparently seeking personal benefits for public work.

"The fact you can easily extend the database functionality was crucial to achieve the result on time."

- *Fabrizio Fortino, Consultant Software Engineer @ RTÉ Investigations Unit.*

One of the team's consultant software engineers, suggested working with a database he had previously used on other projects that might be up to the task. He suggested using the multi-model capabilities of OrientDB, which provided the possibility to connect documents using a graph and to later traverse the documents. *"Another important feature of OrientDB is the ease of use and the extensive support of different drivers."* said Fabrizio Fortino, Consultant Software Engineer for RTÉ Investigations Unit. *"We chose to implement the solution in Groovy with the native Java driver. For some complex operations, we implemented server side functions using the OrientDB Server-Side Functions[1] in both Javascript and Groovy languages. The fact you can easily extend the database functionality was crucial to achieve the result on time."*

When dealing with real-life scenarios, no task comes without challenges. Some of the data sources that needed to be matched lacked common identifiers to link them. In order to resolve this, RTÉ Investigations Unit built a rule-based fuzzy inference engine using the open source project Duke[2].

Fuzzy Inference Engine
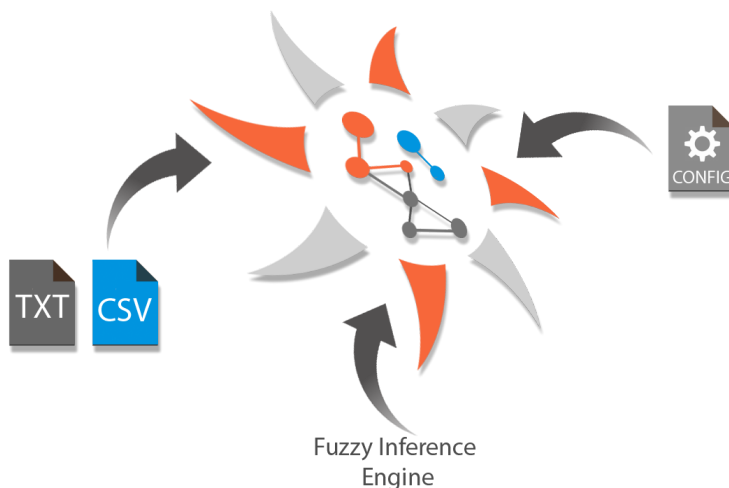
Once the engine was completed, each potential match was then translated into an edge between the detected entities. *"Having the possibility to enrich every edge with the match confidence weight was extremely important in order to easily filter links with a high level of confidence during the data exploration phase."* said Fabrizio Fortino.

"OrientDB had all the features that we needed to tackle the challenge in the agreed time limit. There was no need to learn a new query language since it supports SQL and the Studio Graph View is great for exploratory analysis."

- *Fabrizio Fortino, Consultant Software Engineer @ RTÉ Investigations Unit.*

About RTE

Since its establishment in late 2012, the RTÉ Investigations Unit has produced a number of ground-breaking, award-winning documentaries including "The Torture Files" and "Breach of Trust".

www.rte.ie/news/investigations-unit/

About OrientDB

The native multi-model database combines the connectedness of graphs, the agility of documents and a familiar SQL dialect. OrientDB's graph/document technology has been used by companies such as Nuix and RTÉ investigations Unit for forensic or investigative analysis. Fortune 500 companies, government entities and startups all use the technology to build large-scale innovative applications. Some of their clients include Comcast, Ericsson, the United Nations, Verisign, Pitney Bowes, Sky, Diaku, CenturyLink and Sonatype.

OrientDB recently won the prestigious 2015 Infoworld Bossie award and has been covered by multiple media outlets.

www.orientdb.com

Contact Info

enquiries@orientdb.com

Leveraging on OrientDB's document and graph capabilities enabled RTÉ Investigations Unit to easily and efficiently persist entities as vertices and connect them with edges. To explore the data, they built a series of multiple-step traversal queries, using both SQL and custom functions, with the goal of identifying entities connected by suspect paths not clearly visible in a traditional relational representation. *"OrientDB had all the features that we needed to tackle the challenge in the agreed time limit."* said Fortino. *"There was no need to learn a new query language since it supports SQL and the Studio Graph View is great for exploratory analysis."* Whenever a question arose, they were not only able to rely on the OrientDB team to provide answers, but also found resources and support from the extensive community of developers using OrientDB's open source Community Edition.

Investigating Corruption with Award-Winning Results

The results allowed the Investigations Unit to identify a manageable list of probable undisclosed assets which were used as part of the public interest criteria to set up covert filming. This later exposed an apparent willingness to solicit personal benefits by three elected public representatives.

"Using OrientDB as a heuristic tool, we were able to filter through every elected representative in the country and match them to possible undisclosed assets." said Conor Ryan, Investigative Journalist for RTÉ Investigations Unit. *"The resulting reports identified very clear potential areas of investigation from an otherwise unmanageable and dirty dataset."* Follow up investigations on these assets and the individuals concerned resulted in an **award-winning broadcast documentary**[3] that revealed three elected representatives apparently seeking personal benefits for public work. It also exposed widespread under-declaration of assets by politicians.

Sources:
1 - http://orientdb.com/docs/last/Functions.html
2 - https://github.com/larsga/Duke
3 - http://www.rte.ie/news/investigations-unit/2015/1207/751833-rte-investigates