CASE STUDY

Unlocking the Past: How the Rockefeller Archive Center Digitised Historical Records with Hume

Company:
Rockefeller Archive Center

Business:
Repository and research centre

Challenge:
Produce a Knowledge Graph from unstructured content of Rockefeller's physical paper documents, dating from 1932 to 1941.

Solution:
A knowledge graph powered by Hume and Neo4j.

Results:
Digitalised, structured content, ready for analysis.

Over 10,000 physical typewritten documents from 1932 to 1941 had to be digitised, structured, and connected in order to create a single, centralised source of knowledge, for enabling the analysis of historical processes.

The Rockefeller Archive Center (RAC) is a major repository and research centre for the study of philanthropy and its impact throughout the world. The specific research project revolves around the process of grantmaking, the distribution of funding, the formation of intellectual networks and the transfer and adoption of knowledge.

The RAC wanted the capability to swiftly obtain answers to inquiries such as: "Are there any patterns that typically precede the funding of an idea?" and "Do granted projects tend to run through recommendations of influential scientists or previous grantees?". In order to achieve this, the historical records were brought to life via complex machine learning pipelines, to extract the wealth of knowledge within and modelling it as a knowledge graph.

The project was launched to create a model that could demonstrate how structured data created from analogue, textual records can support a broader range of historical researchers working in quantitative methodologies. With many of the organisations whose records are held by the archive transitioning to paperless and digital platforms, the project was initiated to enhance ways of using historical records.

The Rockefeller Archive Center selected GraphAware as their vendor of choice due to the dynamic capabilities of the Hume knowledge graph platform and the experience of GraphAware data scientists with unstructured data.

The Challenges

As previously mentioned, the project had a set of clearly defined objectives:

- Demonstrate the potential of digitising analogue historical records of various types and combining and connecting them within a single searchable graph database.
- Design advanced Machine Learning pipelines to extract knowledge from the documents to produce a Knowledge Graph.
- Identify new insights and patterns of grantmaking by examining intellectual networks and the development of research fields over time.
- Provide advanced analytics tools to allow researchers from a broad range of quantitative disciplines to engage with the records held by the Rockefeller Archive Center

The primary challenge in this project was to extract structured information from the unstructured content of physical typewritten historical documents. To accomplish this task, a system that integrates graph-based technologies and Natural Language Processing (NLP) had to be designed and implemented. The stages of this process included:

1. Transforming scanned documents into a clean textual content.
2. Extracting the relevant relational knowledge locked within the texts.
3. Representing the extracted knowledge in a rich graph structure.
4. Providing an interface to query, analyse and visualise the highly complex graph, allowing for a comprehensive understanding of the connections between the documents and facilitating new discovery opportunities.

The project focused on two types of records, consisting of a total of approximately 12,000 pages dating from 1932 to 1941.

- The Rockefeller Foundation Officer Diaries are detailed textual records of meetings, calls and various discussions among Rockefeller program officers and scientists, universities and other influencers about ongoing and planned research seeking funding
- The Rockefeller Foundation Board of Directors' Minute Books provide information on the grants that were awarded by the Board, such as the receiving university or organisation, the research topic, and the amount.

The historical nature of the original data, the domain-specific linguistic patterns and the complexity of the relational knowledge contained within all suggest that extracting and modelling all relevant information from these physical documents is highly challenging.

The Solution

To address the challenge, the Data Science Team at GraphAware designed a schema and deployed a solution built on the following pillars:

1. Optical Character Recognition document parsing (OCR)
2. Leverage metadata from OCR to identify
   - Relevant snippets of Board of Directors' Minute Books
   - Individual diary entries from Officer Diaries, any relevant metadata and cleansed text suitable for NLP analysis
3. Customised Named Entity Recognition
4. Coreference resolution for specific historical linguistic patterns
5. Unsupervised entity disambiguation and resolution
6. Entity Relation Extraction
7. Analysis
   - Identification of individual officer-scientist conversations
   - Identification of follow-up conversations
   - Matching Grants (from Board of Directors' books) to Conversations (from Officer Diaries)
   - Graph algorithms/analytics to analyse the intellectual network (identify influencers etc.)
   - Analyse time evolution of research disciplines
   - Configure Hume Actions to allow users to answer relevant questions as specified by RAC

The Knowledge Graph schema was created with the specific requirements of this project in mind, and it was also optimised for maximum user value. The goal of the schema is to provide an intuitive and easy-to-use interface for exploring and answering questions about the data contained in the documents.

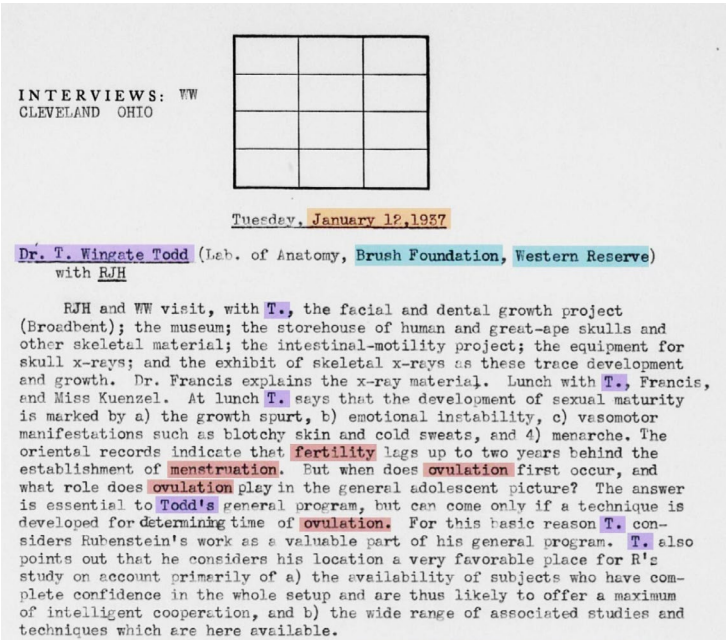
A simplified version of the knowledge graph schema

The Rockefeller Archive Center and GraphAware decided to use the UBIAI annotation platform to label the training dataset using iterative model-assisted labelling technique. The purpose of this was to develop a custom Named Entity Recognition (NER) model that would be able to understand the specific terms and entities present in the records.

Are you curious to discover how advanced analytics & machine learning can be utilised to effectively utilise knowledge?
GraphAware's Lead Data Scientist explains the technologies and complex pipelines employed for this major Knowledge Graph.

The Results

The technical solution, which included OCR parsing, complex customised NLP processing and graph data science to connect and mine the texts from thousands of pages and model them within a knowledge graph, was completed and ready for use within six months. This enabled analysts to effectively answer important questions through the use of graph visualisation, customised by Hume Actions.

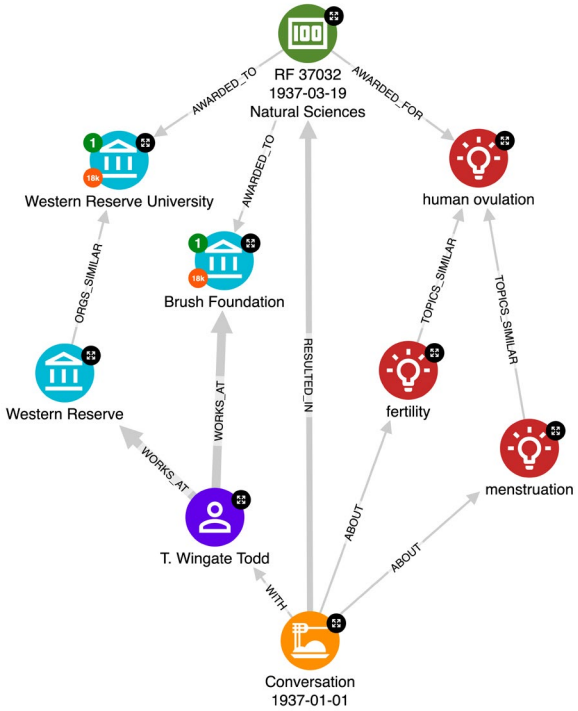


INTERVIEWS: WW
CLEVELAND OHIO

Tuesday, January 12,1937

Dr. T. Wingate Todd (Lab. of Anatomy, Brush Foundation, Western Reserve)
with RJH

RJH and WW visit, with T., the facial and dental growth project (Broadbent); the museum; the storehouse of human and great-ape skulls and other skeletal material; the intestinal-motility project; the equipment for skull x-rays; and the exhibit of skeletal x-rays as these trace development and growth. Dr. Francis explains the x-ray material. Lunch with T., Francis, and Miss Kuenzel. At lunch T. says that the development of sexual maturity is marked by a) the growth spurt, b) emotional instability, c) vasomotor manifestations such as blotchy skin and cold sweats, and 4) menarche. The oriental records indicate that fertility lags up to two years behind the establishment of menstruation. But when does ovulation first occur, and what role does ovulation play in the general adolescent picture? The answer is essential to Todd's general program, but can come only if a technique is developed for determining time of ovulation. For this basic reason T. considers Rubenstein's work as a valuable part of his general program. T. also points out that he considers his location a very favorable place for R's study on account primarily of a) the availability of subjects who have complete confidence in the whole setup and are thus likely to offer a maximum of intelligent cooperation, and b) the wide range of associated studies and techniques which are here available.



Example of the transformation from analogue documents to digital graph

The Actions feature, which includes predefined Cypher queries, allows analysts to easily and quickly retrieve answers to specific questions such as the number of conversations required to secure a grant, key influencers in the grant-making process, and the impact of pre-existing connections within the organisation.

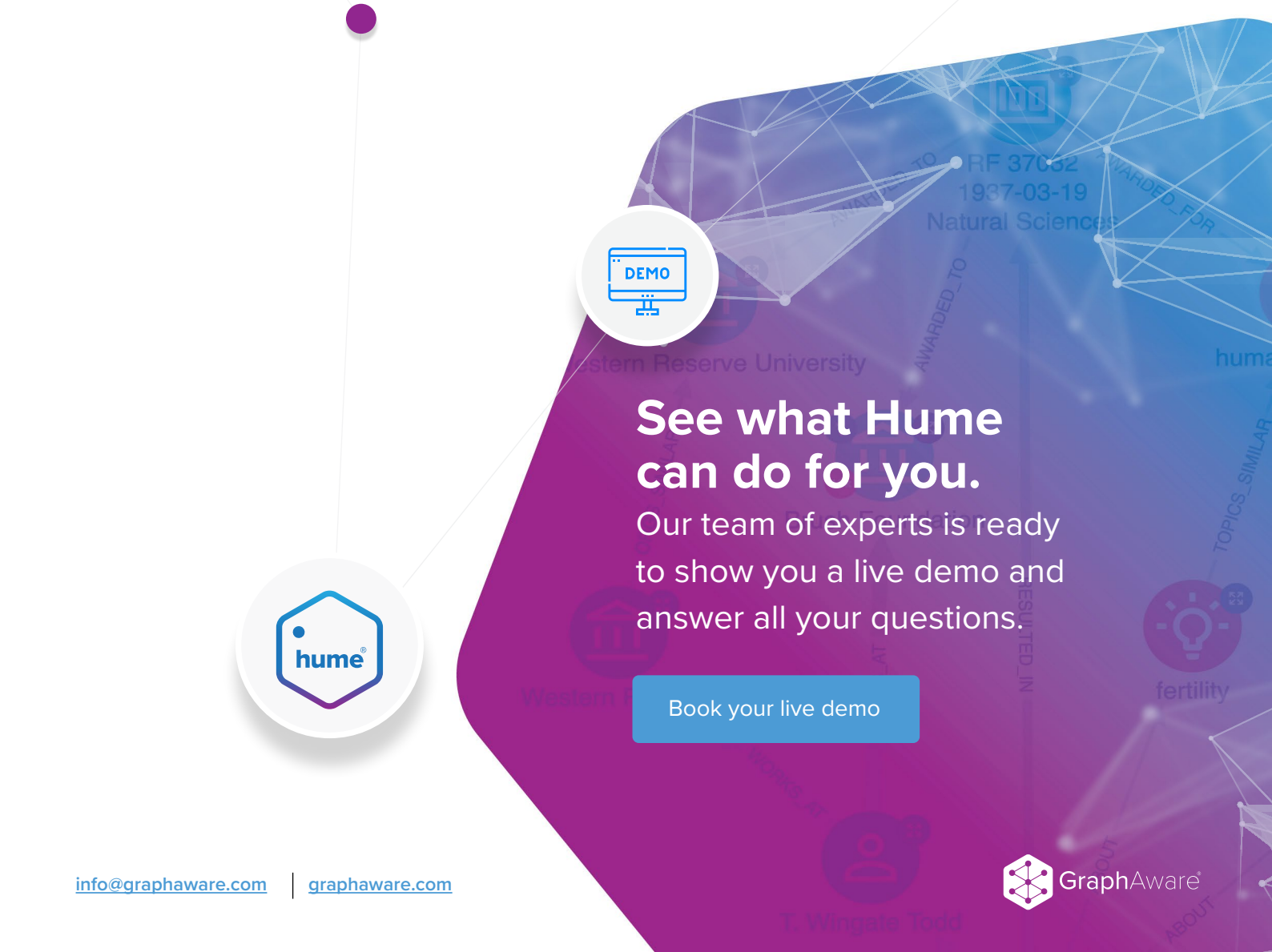
According to the RAC, Hume is a highly advanced and powerful insights and analytics engine; they had not seen anything similar in archival conferences addressing data management. The knowledge graph not only provides insight into past events, such as patterns that led to a grant being approved but also enables users to trace research roads not taken and pathways which did not result in funding.

We are delighted to say that the RAC team stated that, in retrospect, the success of the project would have not been possible without the support and guidance of GraphAware's Data Science Team.

The Future

The success of digitising analogue historical records and creating a queryable knowledge graph using Hume, demonstrated the potential and power of cutting-edge knowledge graph technology. Initially intended as a one-time learning project, the capabilities of the technology impressed the Rockefeller Archive Center's leadership, leading them to acquire a Hume licence and additional services to apply the technology to a second phase of the project.

> "GraphAware is not just a vendor, we gained an intellectual partner"
>
> **Director of Research & Education RAC**



See what Hume can do for you.

Our team of experts is ready to show you a live demo and answer all your questions.

Book your live demo