Final report

Project information and reporting objectives

Project information

Project number:	346068
Project title:	GAIJ - Graph-bound AI Journalism in Financial Fraud
Activity / Programme:	FINANSMARKED
Project manager:	Rydin Gorjão, Leonardo
Project owner:	FAKULTET FOR REALFAG OG TEKNOLOGI, NORGES MILJØ- OG BIOVITENSKAPELIGE UNIVERSITET (NMBU)
Project period:	2024.04.01 - 2025.10.31

Reporting objectives

1.	**Main page of the progress report:** Update progress report up to project completion date.	**Completed**
2.	**Final accounts:** Give a summary of the financial status of the project	**Completed**
3.	**Outcomes and impacts:** I understand that the information entered into the field for Outcomes and impacts will be made publicly accessible*	**Completed**
4.	**Results report:** Attach results report	**Completed**
5.	**Special reports:** Any requests for special reports must be fulfilled. Have special reports been submitted?	**Not applicable**
6.	**Final data management plan:** Has the final data management plan been uploaded?	**Completed**

Final accounts

Actual cost plan (Amount in NOK 1000)

Account	2025	2024	Total sum
Payroll and indirect expenses	887	905	1,792
Procurement of R&D services	0	0	0
Equipment	0	0	0
Other operating expenses	62	0	62
Sum	949	905	1,854

Actual cost code (Amount in NOK 1000)

Account	2025	2024	Total sum
Trade and industry	0	0	0
Research institutes	0	0	0
Universities and university colleges	949	905	1,854
Other sectors	0	0	0
Abroad	0	0	0
Sum	949	905	1,854

Actual funding plan (Amount in NOK 1000)

Account	2025	2024	Total sum
The Research Council	361	1,404	1,765
Own financing	31	58	89
Public funding	0	0	0
Private funding	0	0	0
International funding	0	0	0
Deviation	-557	557	0
Deviation basis	949	905	1,854
Sum	392	1,462	1,854

Comment

The travel budget slightly exceeds the 55kNOK due to the increased costs of travelling. The total costs are in line with the project financing.

Impacts and effects

Anticipated outcomes and impacts - from the grant application form

This project focuses heavily on applications of Artificial Intelligence (AI) tools in Investigative Journalism. GAIJ is a tool that can prove to be of significant importance in classifying and untangling vast amounts of financial data without an extensive need for domain-specific knowledge or the need for extensive examination of the data. A successful implementation of the project can have a long-lasting impact of being the first AI-powered tool that, 1) simultaneously utilises a collection of Large Language Models (LLMs) and 2) use these to classify data, unlike the common application of LLMs that simply generate text. The project is a stepping stone for the development of more complex AI tools that focus on direct, implementable, and practical tools promoting an open and honest society. The project hopes to contribute directly to SDG Target 16.4.

Achieved and potential outcomes and impacts - based on the project results

The project has delivered a functioning prototype that demonstrates, in a bounded way, that Norwegian scanned tax records can be transformed into a structured knowledge graph and explored through an interactive web interface. In practical terms, we have processed on the order of 33 000 documents, constructed a Neo4j graph with companies, people, auditors and addresses, and exposed this via a browser-based tool that a small group of investigative journalists has begun to test. At this stage, however, the actual outcomes are still modest: the system is closer to an advanced demonstrator than to an operational tool that could be relied upon for high-stakes investigations.

The potential impact is considerably larger, but contingent on addressing several non-trivial obstacles. First, the reliance on optical character recognition (OCR) and large-language models (LLMs) means that data quality remains uneven, especially for numerical and tabular content, and systematic validation is currently hampered by the lack of temporally aligned registry data. Second, the computational cost of the extended extraction pipeline severely restricts scalability: extrapolating from current performance, full coverage of the available corpus would require resources well beyond what was available in this project. Third, the graph database and visualisation stack begin to show performance and usability limits as the data size grows, casting doubt on the feasibility of naïvely scaling the current architecture to a truly national scope with substantial IT support.

If these issues can be overcome – through more efficient models, better historical data access, and more scalable graph technologies – the approach has the potential to support more systematic scrutiny of corporate networks and tax practices, and to serve as a template for similar tools in other jurisdictions and domains. At present, however, the impact is best described as laying a technically credible foundation and clarifying the challenges ahead, rather than delivering a ready-to-use instrument for transforming journalistic or auditorial practice.

Results - Summary

Uploaded results - summary

Original filename: GAIJ_progressreport_2025_final.pdf

File reference: RESULTAT_Sluttrapport11904935.pdf

Message to the Research Council of Norway

Special reports

Comment

A full report is included with a denser detail on the project.

Uploaded file

Original filename: GAIJ_progressreport_2025_extended_report.pdf

File reference: SARSKILT_Sluttrapport11904935.pdf

Final data management plan

Uploaded final data management plan

Original filename: DMP_GAIJ.pdf

File reference: DATAHAND_Sluttrapport11904935.pdf

Progress report

Project information and reporting objectives

Project information

Project number:	346068
Project title:	GAIJ - Graph-bound AI Journalism in Financial Fraud
Activity / Programme:	FINANSMARKED
Project manager:	Rydin Gorjão, Leonardo
Project owner:	FAKULTET FOR REALFAG OG TEKNOLOGI, NORGES MILJØ- OG BIOVITENSKAPELIGE UNIVERSITET (NMBU)
Project period:	2024.04.01 - 2025.10.31
Report period:	2024.10.01 - 2025.10.31

Reporting objectives

1.	**Popular science presentation:** I understand that the text of the popular science presentation will be made publicly available*	**Yes**
2.	**Results:** Has information on publications been provided?	**Yes**
3.	**Performance indicators:** All results data that have emerged from the project are to be reported. Has this been done?	**Yes**
4.	**Fellowship grants:** Information regarding all fellowship grants must be complete and correct. Have you updated the man-months and other information for each fellowship-holder?	**Yes**
5.	**International cooperation:** The extent of international cooperation is to be indicated. Has any international cooperation taken place during the report period?	**No**
6.	**Special reports:** If any requests for special reports have been put forth by the case officer at the Research Council, these must be fulfilled.	**No**

Popular science presentation

Popular science presentation (Norwegian)

Store språkmodellar (LLM-ar), som ChatGPT, er i forskingsfronten innan kunstig intelligens. LLM-ar er i stand til å prosessere store mengder tekstdata, inkludert ustrukturert informasjon frå ulike kjelder, slik som nyhendeartiklar, sosiale medium og økonomiske rapportar. Dei kan avdekke mønster i store datasett som eit menneske ikkje greier å sjå, anten på grunn av storleiken på datasettet, eller fordi dei subtile samanhengane mellom ulike element i teksten ikkje er openberre for eit utrent auge. På same måte som eit menneskesinn prøver å finne samanhengar, kan LLM-ar brukast til å teikne eit nettverk eller ein graf av samhandling mellom selskap og andre aktørar. På den måten kan vi prøve å evaluere dei relevante aktørane som er involverte i mistenkelege aktivitetar. Prosjektet freistar å bruke store og små LLM-ar til å lese økonomiske rapportar, skattemeldingar og bankutskrifter. Formålet er å avdekke transaksjonar som er ulovlege eller gjorde med vond hensikt. Det er avgjerande å ta i bruk framveksande teknologiar som LLM-ar for å finne og identifisere illegitime økonomiske aktivitetar.

Popular science presentation - Updated (Norwegian)

GAIJ undersøkte korleis moderne språkteknologi kunne støtte meir systematisk analyse av skattemateriale og tilhøyrande forvaltningsdokument. Som eit pilotprosjekt utvikla prosjektet ein fungerande prototype som vart publisert på nett. Prosjektet hadde som hovudfokus å hente ut strukturert informasjon frå i stor grad ustrukturert tekst, til dømes opplysningar om inntekt, frådrag, eigarskap og omtale av motpartar, og å representere dette i ein konsistent datamodell. Ved å kombinere informasjonsekstrahering med enkel kopling av identitetar og aktørar, bygde arbeidet sporbare samanhengar for “kven–kva–når” som gjorde det lettare for analytikarar å orientere seg i store mengder skattemateriale og peike ut saker som burde undersøkast nærare. Vektlegginga var metodisk og operativ: å styrkje openheit og etterprøvbar tolking av dokument, heller enn å lova automatisk avdekking av lovbrot.

Popular science presentation (English)

Large-language models (LLMs), most famously nowadays, ChatGPT, are at the forefront of interest in research in Artificial Intelligence. LLMs have the ability to process vast amounts of textual data, including unstructured information from diverse sources such as news articles, social media, and, most importantly financial reports. They can unveil patterns in large datasets that are inaccessible to a human reader, either because of the sheer size of the databases or because of the subtle connections between various elements in text that might not be obvious to an uneducated eye. Similar to how a human mind seeks to puzzle evidence together, LLMs can be used to construct a network or graph of interaction between companies and other actors, and therein try to assess the relevant actors immersed in suspicious activities. This project aims to harness the potential of LLMs, big or small, in unveiling links in financial reports, tax declarations, and bank statements, that prove to be illegal or have been made with ill intent. It is crucial to use burgeoning technologies, like LLMs, to aid us in tracking and identifying illegitimate financial activities.

Popular science presentation - Updated (English)

GAIJ explored how modern language technology could support systematic analysis of tax records and related administrative documents. As a pilot project, it produced a working prototype that was published online. The project focused on extracting structured information from largely unstructured text—such as declared incomes, deductions, ownership relations, and references to counterparties—and representing these elements in a consistent data model. By combining information extraction with simple entity linking, the work built traceable “who–what–when” connections that helped analysts navigate large volumes of tax material more efficiently and identify cases that merited closer scrutiny. The emphasis remained methodological and operational: improving transparency and auditability of document interpretation, rather than claiming automated detection of illegality.

Message to the Research Council of Norway

Results

Category: Dissemination

Author(s)	Title	Journal/Publisher /Event	Year	ISSN /ISBN	DOI
Rydin Gorjão, Leonardo; João Machado Lourerio, Nuno; Filipe Pires Alves e Calaim, Nuno; Viswanathan, Pooja; Palumbo, Fabrizio; Krøvel, Roy	GAIJ - Graph-bound AI Journalism	DataSKUP 2024	2024		

Performance indicators

Dissemination measures for the general public

Popular science publications (articles/books, books/articles in the public debate, documents formally circulated for review, exhibitions, fiction, etc..)

2024	2025	Cumulative number
0	0	0

Dissemination measures for users

Reports, memoranda, articles, presentations held at meetings/conferences for project target groups (public sector, trade and industry, organisations)

2024	2025	Cumulative number
1	0	1

Industry-oriented R&D results

New/improved methods/models/prototypes finalised

2024	2025	Cumulative number
0	1	1

Introduction of new/improved methods/models/technology to enhance value creation

Bedrifter utenfor prosjektet som har innført nye/forbedrede metoder/modeller/teknologi

2024	2025	Cumulative number
0	0	0

Companies participating in the project that have introduced new/improved work processes/business models

2024	2025	Cumulative number
0	2	2

Companies participating in the project that have introduced new/improved methods/technology

2024	2025	Cumulative number
0	2	2

New business activity

New companies launched as a result of the project

2024	2025	Cumulative number
0	0	0

New business areas in existing companies, resulting from the project

2024	2025	Cumulative number
0	0	0

Commercial results to which the project has contributed

Licensing agreements signed (excluding software user licenses)

2024	2025	Cumulative number
0	0	0

Patents registered (the same patent in different countries counts as 1 patent)

2024	2025	Cumulative number
0	0	0

New/improved products finalised

2024	2025	Cumulative number
0	1	1

New/improved processes finalised

2024	2025	Cumulative number
0	0	0

New/improved services finalised

2024	2025	Cumulative number
0	0	0

Scientific/scholarly publications

Book/report

2024	2025	Cumulative number
0	0	0

Fellowship grants

Fellowship grants funded under the project

Type of fellowship(s): Visiting researcher grants

Name	Grant period	Fellowship status	Academic degree	Sex
Pooja Viswanathan	2024.04.14 - 2024.10.30	Unchanged	PhD	Female

National identity number	Country of implementation	Country of work	2024
090787	Norway	Portugal	6

Name	Grant period	Fellowship status	Academic degree	Sex
Nuno Filipe Pires Alves e Calaim	2024.04.30 - 2025.10.31	Unchanged	PhD	Male

National identity number	Country of implementation	Country of work	2024	2025
160987	Norway	Portugal	6	10

Name	Grant period	Fellowship status	Academic degree	Sex
Oihane Horno	2024.10.14 - 2025.10.31	Unchanged	PhD	Male

National identity number	Country of implementation	Country of work	2024	2025
270291	Norway	Portugal	1	10

Name	Grant period	Fellowship status	Academic degree	Sex
Nuno Machado Loureiro	2024.08.31 - 2025.05.15	Unchanged	PhD	Male

National identity number	Country of implementation	Country of work	2024	2025
090287	Norway	Portugal	3	4

International cooperation

International cooperation funded under the project (in NOK 1000)

Amount in NOK 1000

Country	2024	2025
Portugal	732	0

Special reports

Comment

A full report is included with a denser detail on the project.

Uploaded file

Original filename: GAIJ_progressreport_2025_extended_report.pdf

File reference: SARSKILT_Framdriftsrapport11904935.pdf

Graph-bound AI Journalism (GAIJ) in Financial Fraud - Progress Report

1. Project objectives and background

The Graph-bound AI Journalism (GAIJ) project set out to develop and demonstrate an open-source, graph-based AI tool that supports investigative journalism on financial crime, with a particular focus on Norwegian company tax records. The core objective has been to transform large volumes of unstructured, scanned financial documents into a structured, queryable knowledge graph that makes complex ownership structures, corporate relationships, and potential fraud indicators visible and analytically tractable.

The project is motivated by the broader challenge of illicit financial flows, which constitute a substantial share of global economic activity and directly undermine efforts to achieve the UN Sustainable Development Goals, notably SDG 16.4 on reducing illicit financial flows. Illicit transactions are deliberately obscured through multi-layered corporate structures, subsidiaries in different jurisdictions, complex contractual arrangements, and opaque accounting practices.

Investigative journalists play a central role in uncovering such practices, but they typically lack the tools and resources to systematically analyse large document corpora and heterogeneous data sources. At the same time, recent developments in Natural Language Processing, and in particular Large Language Models (LLMs), have opened a window of opportunity: these models can parse and interpret complex texts, and open-source variants now make such capabilities accessible beyond proprietary environments.

GAIJ responds directly to this opportunity. The project combines LLM-based information extraction, cross-referencing with official registries, and graph-based data models to construct an interactive platform where journalists can explore company networks, identify “red flags”, and formulate investigative hypotheses. The initial use case has been a large corpus of Norwegian scanned tax documents from 2021, but the underlying architecture is deliberately designed to be modular and extensible to other domains of public-interest data.

2. Results achieved in relation to the objectives

Over the project period, GAIJ has moved from conceptual design to a fully implemented, end-to-end prototype. The principal objective—constructing a modular pipeline that converts unstructured tax records into an interactive knowledge graph—has been achieved.

Concretely, the project has delivered:

- A robust Optical Character Recognition (OCR) pipeline, based on Tesseract, that converts scanned PDF tax documents into machine-readable text at scale.
- A two-stage LLM-based information extraction system: a lightweight model for core company attributes and a more capable model for extended information and risk indicators.
- An automated verification layer that cross-references extracted entities with the official Norwegian business registry (Brønnøysundregistrene) and structures the registry responses via LLM-based parsing.
- A Neo4j knowledge graph that integrates three data sources (tax documents, LLM extraction, and official registry data) into a unified representation of companies, addresses, people, auditors, and their relationships.
- An interactive web-based visualization frontend (JS/D3.js) that exposes this graph to non-technical users through search, filters, and investigative workflows.

In terms of scale, the full pipeline has been run on approximately 33 000 tax documents, which already suffices to stress-test the architecture, populate the graph with a large and heterogeneous set of entities and relations, and demonstrate realistic investigative use cases.

These results align with the initial objectives. The project has shown that complex scanned financial data can be converted into a verified, structured graph; it has also made this graph accessible to journalists and researchers via a functioning, publicly reachable web interface. In doing so, GAIJ provides a concrete proof-of-concept prototype for how AI and graph analytics can be operationalised in investigative journalism on financial fraud.

3. Main R&D tasks and key groups in the implementation

The R&D work has been organised around a sequence of technically distinct but tightly coupled tasks, each corresponding to a module of the pipeline.

1. **Data transformation and OCR**
   The first R&D task was the construction of an OCR pipeline tailored to Norwegian financial documents. This involved evaluating alternative OCR engines, tuning Tesseract for mixed text–table layouts, and automating bulk processing of hundreds of thousands of scanned pages. The outcome is a reproducible workflow that converts each scanned page into plain text ready for downstream analysis.
2. **Cross-referencing with official registries**
   A second key task was the design and implementation of a verification layer that queries the Brønnøysundregistrene APIs for company records. The raw JSON responses are complex and heterogeneous; they are therefore passed through the LLM server, which extracts only the relevant fields and normalises them into a schema suitable for graph ingestion. This step is central to ensuring data quality and to enabling future benchmarking of extraction accuracy.
3. **LLM-based information extraction (simple and extended)**
   A third task involved the development of carefully engineered prompts and extraction schemas for two LLMs deployed on NAIC hardware (see below)

- A smaller Llama 3.2 3B model is used for “simple” extraction (company identifiers, addresses, leadership, subsidiaries) with emphasis on speed and robustness.
- A larger DeepSeek R1 Distill Llama 8B model is used for “extended” extraction, including auditors, document dates, lists of mentioned entities, and a battery of domain-specific “red flags” across financial, transactional, accounting, and liquidity dimensions, co-designed with investigative journalists.

Both phases required iterative refinement of prompts, schema design, error handling, and post-processing to ensure that model outputs are structurally consistent and analytically meaningful.

4. **Knowledge graph construction and integration**
   A fourth core task was the design of the graph model in Neo4j and the development of ingestion routines that merge three distinct JSON sources into a coherent graph. This included defining node types (companies, addresses, people, auditors), relationship types (ownership, location, mention, auditing, parent–child), and node properties (company type, delivery dates, red-flag indicators). The team implemented and tested the full import pipeline on the processed subset of the document corpus.
5. **Interactive visualization frontend**
   Finally, the effort was dedicated to building the JS/D3.js-based frontend that displays the graph to end users. This involved designing search interfaces, filter mechanisms for

investigative workflows, and interactive graph visualizations that remain usable despite high relationship densities. The frontend has been deployed on a dedicated project server and integrated with a Neo4j backend via an API layer.

The project has been led by the AI Journalism Resource Centre (AIJRC), an interdisciplinary research group spanning Oslo Metropolitan University (OsloMet), the Norwegian University of Life Sciences (NMBU), and the University of Agder (UiA).

Crucial domain expertise has been contributed by investigative journalists led by Rune Ytreberg at the Data Journalism Lab (e24 and iTromsø/Polaris Media), who co-designed the red-flag taxonomy and investigative workflows. The Norwegian Artificial Intelligence Cloud (NAIC) has been an essential infrastructure partner, providing and maintaining the servers used for LLM inference, graph storage, and frontend deployment. NFR project 322336.

4. Assessment of implementation and use of resources

Overall, the project's implementation has been efficient and coherent. The modular architecture envisaged in the early design phase has been realised in practice: each pipeline component (OCR, registry integration, LLM extraction, graph construction, visualization) is separately deployable, testable, and replaceable. This modularity removes the risk of future upgrades (e.g. swapping in new models or databases) and is an important long-term asset.

From a resource perspective, the main constraint has been computational capacity for LLM inference. The extended extraction stage typically requires on the order of one minute per document on affordable server hardware, implying several days of continuous computation for tens of thousands of documents. Within the available NAIC resources and project timeframe, the team prioritised depth and reliability of the pipeline over full coverage of the 400 000-document corpus. Approximately 33 000 documents have been fully processed, which is sufficient to validate the architecture and populate a graph large enough for realistic investigations, while providing a clear basis for future scaling with additional resources.

Similar trade-offs arose in the graph database and visualization components. Neo4j and D3.js perform well at the current scale, but the team observed performance degradation and visual clutter as the number of nodes and edges increases. This has been handled through pragmatic design choices (e.g. limiting query scope, tuning visualization parameters) that preserve usability for demonstration and pilot use, while clearly identifying the technical challenges associated with operating at full national scale.

In light of these constraints, the delivered results represent a substantial and judicious use of the allocated resources. The project has produced a working end-to-end system, a publicly accessible demonstration environment, and a clear roadmap for scaling and refinement.

5. Research stays abroad

No formal research stay abroad has been undertaken within the framework of this project.

6. Anticipated significance and benefits of the results

The results of GAIJ have significance at several levels. For the research field, the project contributes a concrete, operational example of how LLMs, official registries, and graph databases can be combined into a coherent architecture for analysing complex financial documents. It advances methodological work at the intersection of AI, data journalism, and network analysis by moving beyond isolated prototype scripts towards an integrated, modular platform. The emphasis on Norwegian-language documents and public registries

also strengthens the evidence base for applying state-of-the-art NLP to mid-sized, non-English data ecosystems.

For the development of expertise, the project strengthens capacity in AI-driven journalism and computational social science within the participating institutions. It has required expertise in NLP, software engineering, graph modelling, and investigative practice, and it has fostered sustained interaction between computer scientists, journalists, and infrastructure providers. This interdisciplinary collaboration is likely to have lasting effects, both in terms of future joint projects and in the training of students and early-career researchers at the partner universities.

For trade and industry, particularly the media sector, the project illustrates how newsrooms and media groups can leverage open-source AI tools and national data infrastructure to support investigative work that would otherwise be infeasible. By lowering the technical barriers to graph-based analysis of financial networks, GAIJ can contribute to more systematic scrutiny of corporate structures, tax practices, and potential fraud.

For society as a whole, the potential benefits lie in increased transparency around corporate behaviour, improved detection of irregular or illegal financial activity, and a stronger public debate on tax justice and regulatory enforcement. In the broader context of the SDGs, the project contributes methodologically to efforts aimed at quantifying and reducing illicit financial flows, thereby supporting more equitable and sustainable economic development.

7. Plans for dissemination and utilisation of the results

The AIJRC plans to disseminate results through a scientific publication and potentially conference contributions in the field of data journalism and AI. The dissemination will primarily cover empirical insights about the structure of using AI in large-scale data with untrained personnel.

Further development of the tool and the utilisation of the results is currently bound to the acquisition of more/new funding to support the scientific staff. The project, as a prototype, shows promise, but a large-scale deployment with real-world, updated data, is still not reachable at this stage.

8. Results expected after project completion

The current prototype marks an important milestone in the interplay between AI tools and large tabular data interaction, but does not exhaust the potential of the GAIJ architecture. Several results could be expected to be finalised after the formal project period.

First, extended network analyses of the constructed graph are possible, in cooperation with partners. These can include the computation of network metrics, detection of structurally interesting subgraphs (e.g. highly connected individuals, addresses associated with many companies), and exploration of patterns associated with specific categories of red flags.

Second, building a consortium which could to explore the application of the pipeline to additional data domains beyond tax records, such as property and zoning documents, public procurement contracts, or other regulatory filings. These extensions will both test the generality of the approach and open new investigative avenues.

Finally, building on the established infrastructure, the next phase will focus more directly on embedding algorithmic fraud detection methods into the graph, moving from rule-based red flags towards pattern recognition and anomaly detection in corporate networks.

All results, the entire setup procedure, and the deployment tool are available on GitHub: https://github.com/AIJRC/GAIJ and abide by an open-source policy. Any interested party can implement GAIJ and use it with their data.

Graph-bound AI Journalism (GAIJ) in Financial Fraud

Comprehensive Progress Report

Abstract

The Graph-bound AI Journalism (GAIJ) project has successfully transitioned from its foundational research and design phase into a period of robust technical execution and integration. Over the past period, the consortium has constructed, tested, and integrated a fully modular pipeline that transforms unstructured Norwegian financial documents into a dynamic, queryable knowledge graph. This report details the significant architectural milestones achieved, including the deployment of a distributed Large Language Model (LLM) inference system for high-accuracy data extraction, the development of a flexible Neo4j graph database backend, and the launch of an interactive web-based visualization frontend. A publicly accessible repository and live demo, showcasing processed data from Norwegian tax records, stands as a testament to the project's viability and provides a crucial platform for stakeholder feedback. While challenges in data heterogeneity and processing scale persist, the project is firmly on track, having built the core infrastructure necessary to advance towards its ultimate goal: deploying AI-powered analytics to detect patterns of financial fraud within complex company networks.

1. Introduction and Project Recap

1.1 The Global Challenge of Illicit Financial Flows

Illicit financial transactions represent a profound threat to global economic stability and equitable development, comprising a staggering portion of the world's economy—estimated at up to 15% of global GDP (Alstadsæter et al., 2018). These illegal flows corrode the foundations of society, directly undermining the achievement of the UN Sustainable Development Goals (SDGs) by widening inequalities, stifling the growth of competitive business, and depleting domestic resources crucial for sovereign security and structural transformation. The UN explicitly recognizes this menace, with SDG indicator 16.4.1 calling for the measurement of the "total value of inward and outward illicit financial flows," and classifying it as a Tier III indicator, highlighting the urgent need for developed methodologies to tackle it (UN Statistics Division, 2020). The central challenge, however, lies in the very nature of these transactions: they are intentionally designed to be hidden, masked by complex layering through subsidiaries, phoney service charges, and country-specific tax loopholes involving licence fees and reimbursements.

1.2 The Role of Investigative Journalism and the AI Opportunity

At the forefront of exposing these malicious activities is investigative journalism, a critical force for accountability. However, the scale and complexity of modern financial data pose immense challenges. News media, including those in Norway, often struggle to cover financial crimes adequately due to the technical expertise and resources required. Technological advancements, particularly in Artificial Intelligence (AI), offer a transformative potential to augment journalistic capabilities. While tools like ChatGPT demonstrate AI's utility for tasks like data scraping and translation, concerns over accuracy and "hallucinations" persist. The International Consortium of Investigative Journalists (ICIJ) has pioneered the use of AI in investigations like the Panama Papers, yet broader application is hindered by high costs, limited training data, and computational complexity. In the near term, the most significant potential for AI in investigative journalism lies in data preparation—specifically, extracting structured information from vast, unstructured document troves and linking records across disparate databases.

1.3 The Emergence of Large Language Models (LLMs)

The most promising AI advancements for this task stem from Natural Language Processing (NLP), which has been revolutionized by the advent of Large Language Models (LLMs). Unlike traditional NLP models trained for narrow tasks, LLMs are general-purpose neural networks with billions of parameters, trained on vast text corpora. Their ability to parse, interpret, and generate human language with remarkable proficiency makes them uniquely suited to handle the complexity and variability of financial documents. The recent proliferation of open-source LLMs (e.g., Llama, BLOOM) has democratized access to this powerful technology, moving it beyond the realm of proprietary systems like BloombergGPT.

1.4 The GAIJ Project

The Graph-bound AI Journalism (GAIJ) pilot project is a direct response to the SDG's call for new methodologies and the transformative potential of AI. GAIJ [pronounced 'gauge'] is an interdisciplinary initiative focused on developing an open-source software tool that precisely

addresses the identified gaps. It brings together graph analysis and large language models to examine irregular transactions between financial entities. The project is founded on the premise that illicit transactions create distinctive, albeit hidden, patterns within complex networks of companies and subsidiaries. By leveraging LLMs to extract structured data from unstructured financial documents and representing it as an interconnected graph, GAIJ aims to make these patterns visible and analyzable.

The project's first phase, as previously reported, was dedicated to confronting the practical challenges of applying this vision to a real-world dataset: a heterogeneous corpus of over 350,000 scanned Norwegian tax documents from 2021-2022. The primary hurdles involved accurate data ingestion via format-specific OCR tools for both text and complex tabular data, and designing a robust strategy for data harmonization using a graph-based database model (Neo4j). This current report outlines the successful transition from that design phase to the implementation and integration of a fully functional, modular pipeline, representing a significant milestone in creating a practical tool for investigative journalism and financial transparency.

2. Technical description: Pipeline modular architecture

The core achievement of this period has been the construction and integration of a sophisticated, multi-stage pipeline. Each component is designed to be modular, allowing for independent development, testing, and replacement as technologies evolve.
The GAIJ project transforms complex, unstructured tax records into a powerful investigative tool through a five-stage processing pipeline:

1. **OCR & Data Transformation**
   Scanned tax documents are converted from image-based PDFs into machine-readable text using a custom Optical Character Recognition (OCR) pipeline. This step enables downstream language models to “read” and analyze the content.

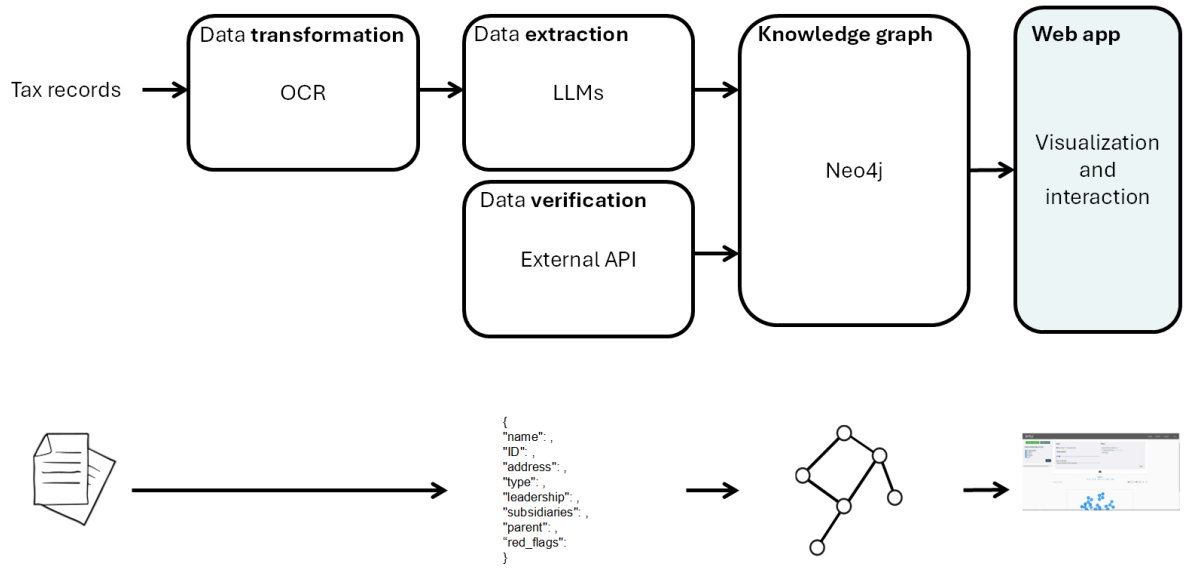
2. **Data Verification**
   Extracted information — such as company names, addresses, and identification numbers — is cross-referenced with authoritative data from Norway’s official business registry (Brønnøysundregistrene). This ensures accuracy, builds trust, and provides journalists with verified company details directly within the system.

3. **Information Extraction with LLMs**
   Large Language Models (LLMs) parse the verified text to extract key structured data and perform deeper analysis to identify additional entities, relationships, and potential “red flags” linked to financial, accounting, or transactional risks.

4. **Knowledge Graph Construction**
   Verified data from official sources and LLM outputs are combined into a **graph database**, linking companies, people, auditors, and addresses. This network view reveals hidden relationships and patterns that are difficult to detect from documents alone.

5. **Interactive Visualization Platform**
   The structured graph data is made accessible through an intuitive web interface. Users can explore connections, trace ownership structures, and investigate risk indicators through queries and interactive visualizations.

GAIJ modular architecture



2.1 Data transformation

A significant challenge in the GAIJ project has been the initial data transformation phase: converting the raw input— scanned tax document images—into a machine-readable text format suitable for analysis by Large Language Models (LLMs).

Our primary data source consists of PDF scans of financial declarations. For an LLM to extract structured information (like company names or ownership details), it must first be able to "read" the text within these documents.

To solve this, we implemented an Optical Character Recognition (OCR) pipeline. OCR technology analyzes the shapes of characters in an image and converts them into digital text. After a rigorous evaluation of several available OCR libraries, we determined that the Tesseract engine (accessed via the tesserocr Python library) provided the best combination of accuracy, speed, and reliability for our specific dataset, which contains a mix of typed text and tabular data in Norwegian.

We successfully developed and automated a robust OCR pipeline that processes all document images in bulk. This system efficiently transforms each scanned page into a plain text file, creating the essential raw material for the next critical stage of our project: Information Extraction using LLMs.

2.2 Cross-referencing with official sources

A critical objective of the GAIJ project is to ensure the accuracy and reliability of the information extracted from tax documents. To address this, we implemented a data verification mechanism that cross-references our LLM-extracted data with official, authoritative sources.

An important portion of the data found in the tax documents (such as company names, addresses, and organization numbers) is also publicly available through an API provided by

the Brønnøysundregistrene[1], Norway's official registry for business entities. This provided a golden opportunity to validate our results against a "ground truth" source.

We developed a dedicated pipeline that automatically retrieves and processes this official data. This serves two key purposes:

1. For Developers: It allows us to quantitatively measure the accuracy of our primary data extraction pipeline.

2. For End-Users: It provides a means for journalists using the tool to verify key company details directly within the system.

The pipeline operates through the following automated steps:

1. **API Call:** For a given company tax ID (organisasjonsnummer), the system calls the relevant Brønnøysundregistrene API endpoints to fetch the official, up-to-date company record.
2. **Data Extraction:** The raw JSON response from the API can be complex and contain superfluous information. To distill it into the clean, structured format required by our graph database, we send this raw data to our LLM server (hosting the Llama model). The LLM is prompted to act as a sophisticated data parser, extracting and translating only the relevant fields (e.g., company name, address, board members) into a predefined schema.
3. **Structured Data Output:** The processed and cleaned data is received from the server and saved in a standardized JSON format. This creates a verified dataset that is consistent with the data extracted directly from the tax documents.4

This verification pipeline is fundamental to the project's integrity. It not only enables us to benchmark and improve our core OCR/LLM extraction accuracy but also equips end-users with a powerful feature to confirm critical company information, thereby increasing trust and utility in the GAIJ tool for investigative journalism.

2.3 Data Extraction

A central component of the GAIJ project involves the extraction of structured and meaningful information from complex tax records. To achieve this objective, we used the previously converted text form of the data and used Large Language Models (LLMs) with carefully designed prompts to automate and standardize the information extraction process.

The overall data extraction workflow was implemented in two main stages: a simple data extraction phase aimed at capturing fundamental company details, and an extended data extraction phase designed to retrieve more complex information and assess potential risks within the documents.

Simple data extraction

The initial stage of the extraction process focused on the retrieval of essential company-related information using a lightweight and high-performance LLM, *Llama 3.2 3B Instruct*. The primary objective of this phase was to identify and extract key data fields from the text, including: **Company name, Company identification number (ID), Address,**

[1] https://www.brreg.no/

Company type, Leadership information, and Subsidiaries.

This process was executed by deploying the model on a dedicated server and accessing it through API calls. Each request included the tax record text together with a structured prompt that specified the desired output format. The LLM was instructed to return the extracted information in JSON format, following a predefined schema with standardized field names. Once the model generated its output, the data underwent a post-processing stage to verify its accuracy, ensure structural consistency, and confirm compliance with the expected schema. After this validation step, the finalized information was stored in JSON files, providing a structured and reliable foundation for subsequent analytical tasks.

Extended Data Extraction

The second stage aimed to enrich the dataset with additional information and conduct a preliminary risk assessment of the tax records. For this purpose, a more advanced and computationally intensive LLM, *DeepSeek R1 Distill Llama 8B*, was employed due to its superior capability in handling complex and context-dependent extraction tasks.

In this phase, the model was instructed to identify and extract the following supplementary data elements: **Auditor name, Tax document delivery date, All companies mentioned in the document, and all individuals mentioned in the document.**

This information enhanced the basic dataset produced during the initial extraction phase, enabling a deeper and more comprehensive analysis of the tax records.

In addition, guided by the expertise of investigative journalists specializing in tax record analysis, the LLM was tasked with identifying and flagging potential **risk indicators** (referred to as *red flags*). These were categorized as follows:

Financial Red Flags:
- Complex financial instruments (e.g., derivatives, structured products, swaps with unclear terms)
- Undisclosed operating lease obligations
- Guarantees or pledges insufficiently documented
- Asset write-downs (e.g., impairments, market declines)
- Reliance on external financing for core operations

Transactional Red Flags:
- Non-recurring or extraordinary transactions
- Inter-company transactions
- Large receivables with significant collection risks

Accounting Red Flags:
- Auditor qualifications or expressed concerns
- Changes to accounting policies
- Adjustments to prior-year accounts
- Large deferred tax assets with uncertain realization
- Abnormal tax payments or disputes
- Unaudited financial statements
- Pending litigation or legal claims

Liquidity Red Flags:

- Negative cash flow despite reported profitability
- Significant pension obligations

For each red flag identified, the model produced a Boolean assessment (True/False) indicating its presence and provided the specific location within the document that supported its conclusion.

In addition, a simple search was employed to flag the usage of a **list of keywords** ("kompensasjon", "sluttavtale", "oppsigelsesdato", "oppsigelse", "sluttdato", "opphør", "trukket", "etterlønn", "bonus", "variabel lønn", "resultatbasert", "milepæl", "etterbetaling", "etterbetalt", "privatlån", "private lån", "selgerkreditt", "interntransaksjon", "diskresjonær", "låneforfall", "forfalt", "ubetalt", "solgt aksjer","covid","covid-19", "Kjell Inge Røkke")

linked with tax fraud.

The extended extraction phase followed a similar implementation approach to the initial stage. It was conducted through server-based API calls, and the resulting outputs underwent thorough post-processing and normalization to ensure compatibility with the project's standardized data schema. The finalized data was then stored in JSON format for integration into the broader analytical pipeline.

2.4. Knowledge Graph Construction

The second core pillar of the GAIJ project focuses on the construction of a relational knowledge graph to represent and explore the complex relationships contained within the extracted tax data. While the raw information obtained from the documents is valuable on its own, its full analytical potential emerges when it is structured into a graph database, enabling the visualization and querying of intricate networks that are otherwise difficult to detect.

To achieve this, we integrated data from three distinct JSON sources: (1) the company data from the Brønnøysundregistrene API, (2) the core entity information from the Llama 3.2 model, and (3) the in-depth analysis and red flags from the DeepSeek model. These datasets were combined and imported into a Neo4j graph database, which was chosen for its powerful graph-oriented data modeling capabilities and flexibility in handling complex relationships.

Within the graph, information was organized into specific **nodes** and **relationships**, forming a structured representation of the entities and their interactions. The primary nodes created were:

- **Company**: Individual companies.
- **Address**: Registered locations associated with companies.
- **People**: Individuals linked to the companies, such as executives or board members, or people that are mentioned
- **Auditor**: Auditing entities identified in the tax documents.

The relationships between these nodes captured the nature of their connections and included:

- **Located at**: Linking a company to its registered address.
- **Owned by**: Indicating ownership relationships between companies or between individuals and companies.

- **Child of**: Representing subsidiary or hierarchical corporate structures.
- **Mentioned**: Capturing references to entities or people within the documents.
- **Auditor**: Linking companies to their corresponding auditors.

In addition to these structural relationships, company nodes were enriched with several key properties, including their **unique identifier (ID)**, **company type**, **delivery data** and any **red flags** identified during the extended extraction phase. This metadata provided an added layer of analytical depth, allowing for more targeted exploration of potential risks or anomalies.

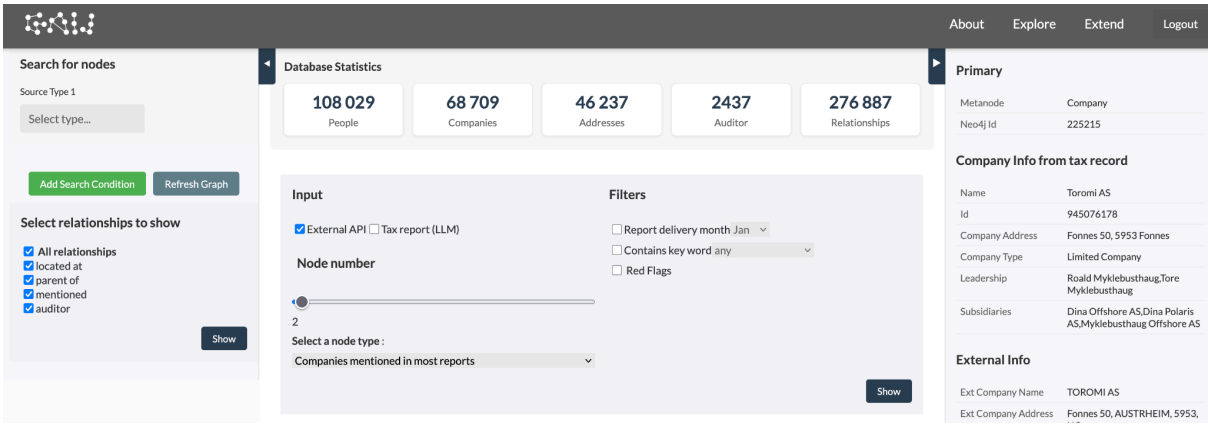


Figure 2. A view from the webapp showing basic statistics about the Graph, as well as sample information from one company.

By structuring the extracted information in this way, the knowledge graph enabled the identification and visualization of networks of companies, individuals, and auditors that were not immediately evident from the tax records alone. This relational perspective greatly enhanced investigative and analytical capabilities, allowing users to trace connections, detect hidden relationships, and uncover patterns that could inform further scrutiny or journalistic inquiry. In addition, this tool allows for the user to interact with the data in a more sophisticated manner than through the webpage, as it can handle complex and unique queries that satisfy the specific user needs.

2.5. Interactive Visualization Frontend

The final component of the GAIJ project pipeline is the development of an interactive web-based visualization platform designed to make the knowledge graph accessible and actionable for non-technical users, such as investigative journalists, researchers, and policy analysts. This interface transforms the underlying graph database into an intuitive, user-friendly tool that supports both exploratory analysis and targeted investigations.

The visualization platform was developed using a combination of JavaScript (JS) and the D3.js visualization library, adhering closely to the architecture proposed in the initial project design. This technology stack was chosen for its flexibility, performance, and suitability for building responsive, data-driven user interfaces. The frontend connects directly to the Neo4j graph database through an API layer, enabling real-time queries and dynamic data rendering in the browser.

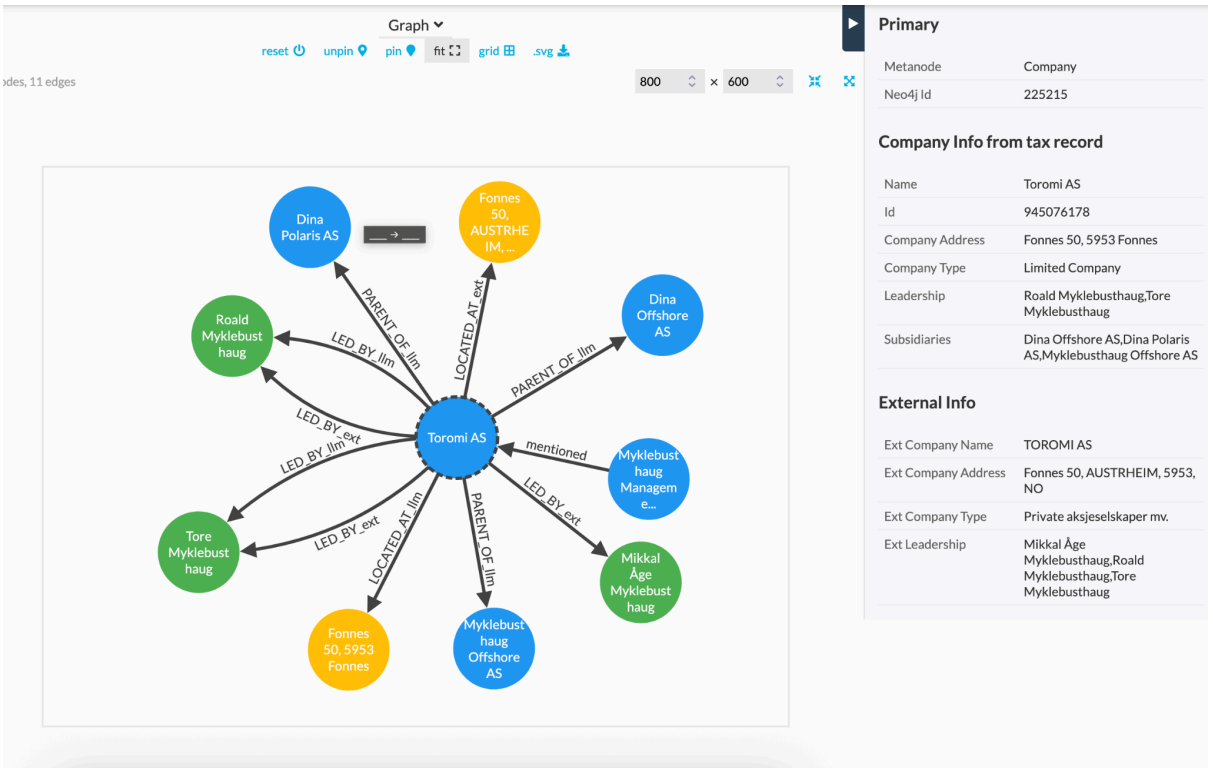


Figure 3. Graph visualization for a sample company. Different entity types are displayed with different colors. Furthermore, the relationship type is also displayed on top of the relationship edge.

3. Functionality

GAIJ project is a web-based platform that allows the user to extract information about Norwegian companies. The platform offers several modes of interaction with the data.

The visualization frontend is deployed on a dedicated project server and is accessible at http://158.37.66.6:8765. This live demonstration environment serves as a practical showcase of the project's capabilities, enabling partners, stakeholders, and investigative teams to directly interact with the knowledge graph and evaluate its potential applications in real-world tax fraud investigations.

Overall, this visualization tool provides a unique and accessible interface for non-technical users to engage with complex financial data. By lowering the barrier to advanced graph-based analysis, it supports more effective tax fraud detection and investigation.

3.1. Searching function

The most straightforward approach enables users to search directly for specific entities — including companies, individuals, or addresses — and immediately visualize their connections within the broader network. Once a user selects a node (for example, a company), the application displays its associated properties, including metadata such as company type, ID, and any red flags identified during the data extraction process. This feature allows users to quickly understand the relationships and potential risk indicators linked to a particular entity.

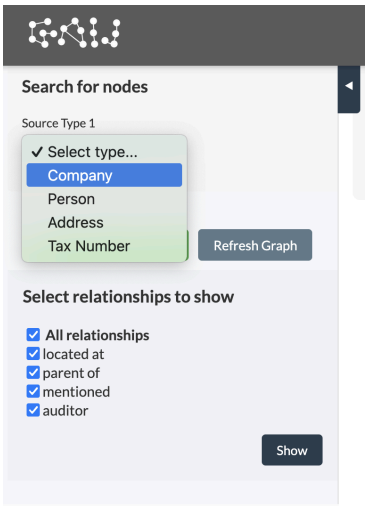


Figure 4. Search fields available to query the Graph Network

3.2. Search based on filters

Beyond simple search and navigation, the platform also provides a powerful filtering and discovery system that supports more advanced investigative workflows. This feature enables users to dynamically generate visualizations based on specific analytical criteria, helping them uncover potentially suspicious patterns that may not be immediately visible. Users can, for example, filter and display:

- Companies with a high number of subsidiaries or board members.
- Companies mentioned in the greatest number of tax reports.
- Addresses or auditors associated with multiple companies.
- Individuals who serve as leaders in numerous companies.
- People frequently mentioned across corporate records.
- Cases where a parent and subsidiary share the same leadership.

Additional filtering options further refine these analyses. Users can limit results to companies whose tax documents were submitted within a specific month, as late filings may indicate suspicious activity. It is also possible to filter based on keyword presence in reports or restrict the visualization to entities associated with identified red flags.

Once a filtered graph is generated, users can interactively explore the resulting network, examining the connections and relationships revealed by their chosen parameters. This capability significantly enhances the investigative process, allowing for deeper insights into complex corporate structures, potential conflicts of interest, and hidden relationships that could indicate fraudulent or high-risk activity.

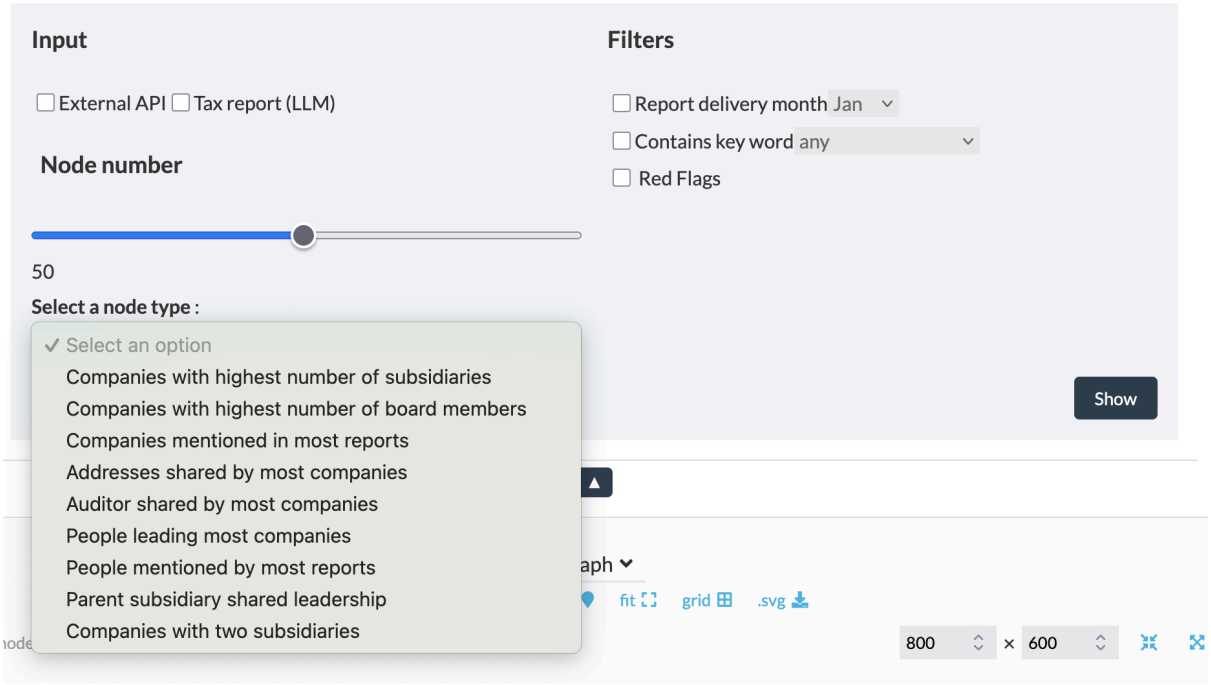


Figure 5. Investigative workflows available to the user. These fields can be customized by the developers of the platform to aid the users with any research question they have.

4. Contributors

4.1 Project managers and groups

This Project has been led by the AI Journalism Resource Centre (AIJRC), the research group founded in 2022 at the Oslo Metropolitan University (OsloMet), which now extends to the Norwegian University of Life Sciences (NMBU) and the University of Agder (UiA). The research team is led by Roy Krøvel (Professor in Journalism at OsloMet) and Fabrizio Palumbo (Associate Professor in Journalism AI at OsloMet/UiA).

The AIJRC also benefits from robust institutional support from OsloMet, NMBU, and UiA, which provide essential resources such as IT infrastructure, access to data, library services, and specialized equipment. Together, these universities not only sustain the group's research and development efforts but also play an active role in educating the next generation of data scientists and journalists in the rapidly evolving field of AI-driven journalism.

4.2 Researchers

This work has been done by Nuno Calaim, Nuno Loureiro, Pooja Viswanathan, Oihane Horno and led by Leonardo Rydin Gorjão and Fabrizio Palumbo. The management of the group was carried out by Fabrizio Palumbo and Roy Krøvel.

4.3 Collaborations

This project was made possible through a strong interdisciplinary collaboration between our research team, a group of investigative journalists, and the Norwegian Artificial Intelligence Cloud (NAIC) infrastructure. Each partner contributed unique expertise that was essential to the project's success.

Investigative Journalism Partners

To ensure that our tool addressed real-world needs in the context of tax fraud investigations, we collaborated closely with a team of investigative journalists throughout the project. Through regular meetings and workshops, they provided crucial insights into how journalists approach complex financial investigations, helping us define the most important functionalities our system should offer. They also guided us in identifying key “red flags” commonly used in the field to detect suspicious activity.

The team was led by Rune Ytreberg, editor at the Data Journalism Lab at the daily newspaper *iTromsø*. The lab leverages AI technologies to develop innovative editorial tools for over 70 newspapers within Polaris Media, one of Scandinavia’s largest media groups. Rune holds a master’s degree in business journalism and has received multiple awards for his investigative reporting. He was awarded the prestigious SKUP Award for the NRK Brennpunkt documentary *Rovfiskerne*, which uncovered major environmental crimes in the Barents Sea.

Rune also leads several investigative projects funded by Fritt Ord, the Norwegian Research Council, and Skattefunn, in collaboration with researchers from OsloMet, UiB, NMBU, UiT, IBM, and NORCE. These projects employ NLP and AI to investigate potentially illegal or unsustainable economic activities, including *Byenvår*, *Eiendomsrobot*, *Byggebot*, and *Fishdata*. Over the past two decades, he has lectured extensively on investigative and data journalism at universities and conferences across Europe.

Norwegian Artificial Intelligence Cloud (NAIC)

The Norwegian Artificial Intelligence Cloud (NAIC) was a crucial technical partner in this project. Their infrastructure enabled us to deploy and scale the computational resources needed for our work. NAIC supported the setup and maintenance of the servers that hosted our analysis pipelines, the graph database, and the interactive web platform. Their support was instrumental in ensuring the system’s robustness, scalability, and performance throughout the project lifecycle. NFR project 322336.

4. Collaborations and outreach

We promoted our tool at DataSkup, a Norwegian journalism conference, held on the 26th of October, 2024 at OsloMet, Oslo, https://www.journalismfund.eu/data-skup-2024. We have similarly interacted with e24, iTrømso, and NRK, within the project, in various online meetings to showcase our tool. We have also contacted and discussed the project with the Brønnøysund Register (https://www.brreg.no/en/), the national body of Norway.

5. Constraints and limitations

While the GAIJ project has successfully demonstrated a novel approach to extracting and structuring information from scanned tax documents, several important constraints and technical limitations influenced the scope, performance, and outcomes of the work. These challenges arose at multiple stages of the pipeline; from data ingestion and preprocessing to information extraction, storage, and visualization, and offer valuable insights for future development and scaling.

The main limitations encountered can be grouped into three key areas: (1) challenges related to Optical Character Recognition (OCR) and the nature of the source data, (2)

processing time and scalability constraints associated with large language model (LLM) analysis, and (3) performance issues in managing and visualizing large-scale graph databases.

5.1 Optical Character Recognition

One of the most fundamental challenges of the project stemmed from the nature of the original data: all tax records were provided as scanned images of printed documents. This presented a significant technical hurdle, as image-based data must first be converted into machine-readable text before any meaningful analysis can be performed.

Implementing an Optical Character Recognition step addressed this challenge, but it also introduced new complexities and limitations:

- **Increased processing time:** Each document required an additional preprocessing step, extending the overall time needed to process large datasets.
- **Loss of numerical fidelity:** While OCR performed reliably for textual data, it struggled with numerical values, tables, and financial figures, which are central to tax records. This limitation meant that direct verification or analysis of reported numbers was not feasible within the current pipeline.
- **Scope restriction:** Because of these technical constraints, the project's analytical focus had to be narrowed to textual and relational information, leaving out quantitative aspects such as financial consistency checks or numerical anomaly detection.

5.2 Analysis time

Another significant limitation encountered during the project was the time required for LLM-based information extraction, which became a major bottleneck when scaling the analysis to large document collections.

The two-stage extraction process varied considerably in terms of computational demand:

- **Simple Extraction:** The initial phase, focused on identifying fundamental company details, required approximately 3–6 seconds per document, which was acceptable for most use cases.
- **Extended Extraction:** The second phase, which included red flag assessment and extraction of more complex fields, was far more computationally intensive — averaging 60–70 seconds per document, depending on document length and complexity.

As a result, the total processing time averaged 65–75 seconds per document. On standard, affordable server hardware (which included limited RAM), this posed a considerable scalability challenge. Processing 10,000 documents required approximately 7.5 days of continuous computation. Given a dataset on the order of 400,000 documents, the time requirement would be prohibitive without substantial hardware upgrades or architectural changes.

Due to these constraints, the project team was able to process only about 33,000 documents — roughly one-tenth of the available dataset. This limitation illustrates the need for either more efficient model architectures, distributed computing solutions, or pre-filtering strategies

to prioritize the most relevant documents for deep analysis in future iterations.

5.3 Graph Database Performance

The final stage of the pipeline, constructing and visualizing a knowledge graph, also presented significant scalability challenges. As the number of processed documents increased, so did the number of nodes (representing entities such as companies, people, and auditors) and edges (representing relationships such as ownership or mentions).

This growth, particularly in the number of relationships, followed an exponential pattern, which affected both database performance and the user experience in visualization tools:

- **Performance degradation:** Query response times increased as the database grew, especially for complex queries involving multiple relationship types.
- **Visualization complexity:** The sheer number of nodes and edges made interactive visualization more challenging, with cluttered network graphs and slower rendering speeds.

To partially address these issues, several parameters and visualization strategies were adjusted, which improved usability in smaller and medium-sized datasets. However, it remains uncertain how the system will perform when the full dataset (300,000+ documents) is processed and ingested, as this would represent a significant stress test for both the Neo4j database and the D3.js-based visualization frontend.

Future work will likely require optimization of the data model, partitioning strategies, or the integration of more advanced graph visualization libraries designed to handle large-scale, high-density networks.

6. Next Steps and Future Direction

The GAIJ project has demonstrated the feasibility and potential of using advanced language models, knowledge graphs, and visualization tools to support investigative journalism on complex financial documents. However, the current implementation represents only the first stage of a broader research and development trajectory. Several areas of future work could significantly expand the system's analytical capabilities, improve its reliability, enhance user interactivity, and extend its applicability to other domains. The following subsections outline key directions for future development and research.

6.1 Validation and benchmarking

A critical next step is the rigorous validation and benchmarking of the information extraction pipeline. One of the main motivations for integrating official company data from the Brønnøysundregistrene registry was to establish a reliable reference dataset against which the performance of the LLM-based extraction system could be measured. However, the current situation presents a challenge: the registry data available to the project corresponds to the year 2024, while the scanned tax documents in our dataset are from 2020. As a result, any discrepancies identified between the two datasets cannot be confidently attributed to model errors, as they may instead reflect legitimate changes in company status over the intervening years.

To address this, it will be essential to obtain historical datasets from Brønnøysundregistrene that match the time period of the tax documents under analysis. Access to such data would

enable precise error analysis, performance benchmarking, and systematic improvement of extraction models.

6.2 Network analysis

While the current analytical framework allows users to query the knowledge graph and retrieve structured insights, there is significant untapped potential in applying network analysis techniques to the extracted data. Moving beyond simple queries, future work could focus on identifying and analyzing complex relational patterns within the graph.

For instance, network metrics could be used to detect clusters of companies sharing common ownership structures, reveal addresses linked to numerous corporate entities, or identify key individuals who act as central nodes in corporate networks. Such analyses could uncover subtle or hidden connections that are not immediately visible through standard querying, providing powerful investigative leads and enhancing the system's value as a journalistic tool.

6.3 Interactive data extraction

Another major area for future development involves enhancing user interactivity: both in how data is extracted and how it is queried. This would bring the platform closer to a true human-in-the-loop investigative tool.

From the perspective of **LLM interaction**, one envisioned feature is the ability for users to design and submit their own prompts to extract specific types of information not originally included in the system's schema. The results of such custom extractions could then be incorporated into the knowledge graph, enriching the database for all users. Realizing this feature, however, introduces several layers of complexity. It would require secure and dynamic access to the analysis server, robust post-processing mechanisms to ensure that user-generated prompts yield high-quality data, and regular quality control. potentially through a combination of automated validation and expert review. to prevent data degradation.

On the **graph query** side, a complementary enhancement would be to allow users to write and execute their own queries directly against the Neo4j database. This would provide much greater flexibility in how the graph is explored and visualized, enabling users to construct highly specific investigative questions. However, this approach would require users to possess a working knowledge of graph query languages (such as Cypher), which may limit accessibility. Future work might therefore focus on designing more intuitive query interfaces that translate natural language queries into formal graph queries automatically.

6.4 Interactive data addition

Another promising direction involves enabling users to upload and process their own document collections through the pipeline. Such functionality would provide greater flexibility and customization, allowing journalists, researchers, or organizations to analyze proprietary or confidential datasets without compromising privacy.

This feature could transform the system from a fixed investigative platform into a general-purpose analytical service, broadening its applicability beyond the specific tax datasets used in the current project. It would also open up opportunities for collaborative investigations, where multiple stakeholders contribute data and share insights through a common analytical framework.

6.5 Expansion to New Data Domains

Finally, the versatility of the GAIJ pipeline makes it well-suited for adaptation to a wide range of text-based datasets beyond tax documents. Future work could explore its application in fields such as urban planning (e.g., zoning and property records), academic research (e.g., scientific literature and funding data), or public procurement and policy (e.g., government contracts and spending reports).

Applying the system to new types of text records would not only validate its robustness and generalizability but could also uncover new types of relationships and investigative opportunities. Testing the pipeline in these diverse contexts would further refine its capabilities and contribute to the development of a more flexible, domain-agnostic investigative platform.

6. Conclusion

The GAIJ project has successfully overcome its initial technical hurdles and delivered on its promise of a functional, modular, and open-source pipeline for transforming financial documents into an interactive knowledge graph. The existence of a live demo marks a pivotal moment, moving the project from theoretical research into a phase of practical application and iterative refinement. The team remains focused on enhancing the system's accuracy and scalability, and is now poised to begin its most ambitious work: embedding AI-driven fraud detection directly into the graph it has built, thereby creating a tool for financial transparency and investigative journalism.

All results, the entire setup procedure, and the deployment tool are available on GitHub: https://github.com/AIJRC/GAIJ and abide by an open source policy. Any interested party can implement GAIJ and use it with their data.