

# Fujitsu RDF Use Cases and benchmarking requirements

3rd LDBC TUC Meeting  
London, UK

Nuno Carvalho  
Senior Researcher  
Fujitsu Laboratories of Europe Ltd

[Nuno.Carvalho@uk.fujitsu.com](mailto:Nuno.Carvalho@uk.fujitsu.com)

## ■ Use cases

- Anomaly detection on public sector
- Healthcare
- Financial

## ■ Limitations and benchmarking requirements

# ANOMALY DETECTION

- Generate public sector Linked Data for detecting anomalies and identify cases submitting fraudulent claims
  - Linking data from different councils
  - Linking blacklist of fraudsters
- Effect of using Linked Data
  - use fraud prevention approaches by using a variety of criteria (including credit histories).
  - aggregate information in order to conduct further analysis and investigate allegations of benefit fraud.
- Technical features
  - Technology for identifying fraudsters as early as possible
  - Technology for linking different types of public data

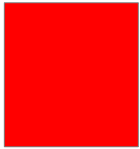
# Anomaly detection use case

## Council Data

Hounslow



Ealing



Hillingdon



**Blacklist of  
pre-identified  
fraudsters**

## Linked Data

Linking Data from different councils together and reconciliation against Blacklist of fraudsters



Anomalies Detection to identify potential frauds

## DWP Staff



Identify potential cases of fraud and help DWP fraud inspectors to reduce their time on looking into a large quantity of data and cross-checking against a blacklist of fraudsters to gather further evidence

## ■ Import data from heterogeneous formats

- Each council has their own data silo, sometimes in formats such as Excel
- The “schema” used in each silo does not differ much, but it needs to be mapped to a common format

## ■ Data analytics

- Query and analyse existing claims from different councils, to find co-relations and frauds

## ■ Incremental processing

- Avoid run analytics on all claims frequently, instead run incrementally when a new claim is generated

# HEALTHCARE

## ■ Integrating Clinical Trials Linked Data with publications

- Linking Clinical Trials
- Linking Scientific Publications
- Linking related twitter or social media

## ■ Effect of using Linked Data

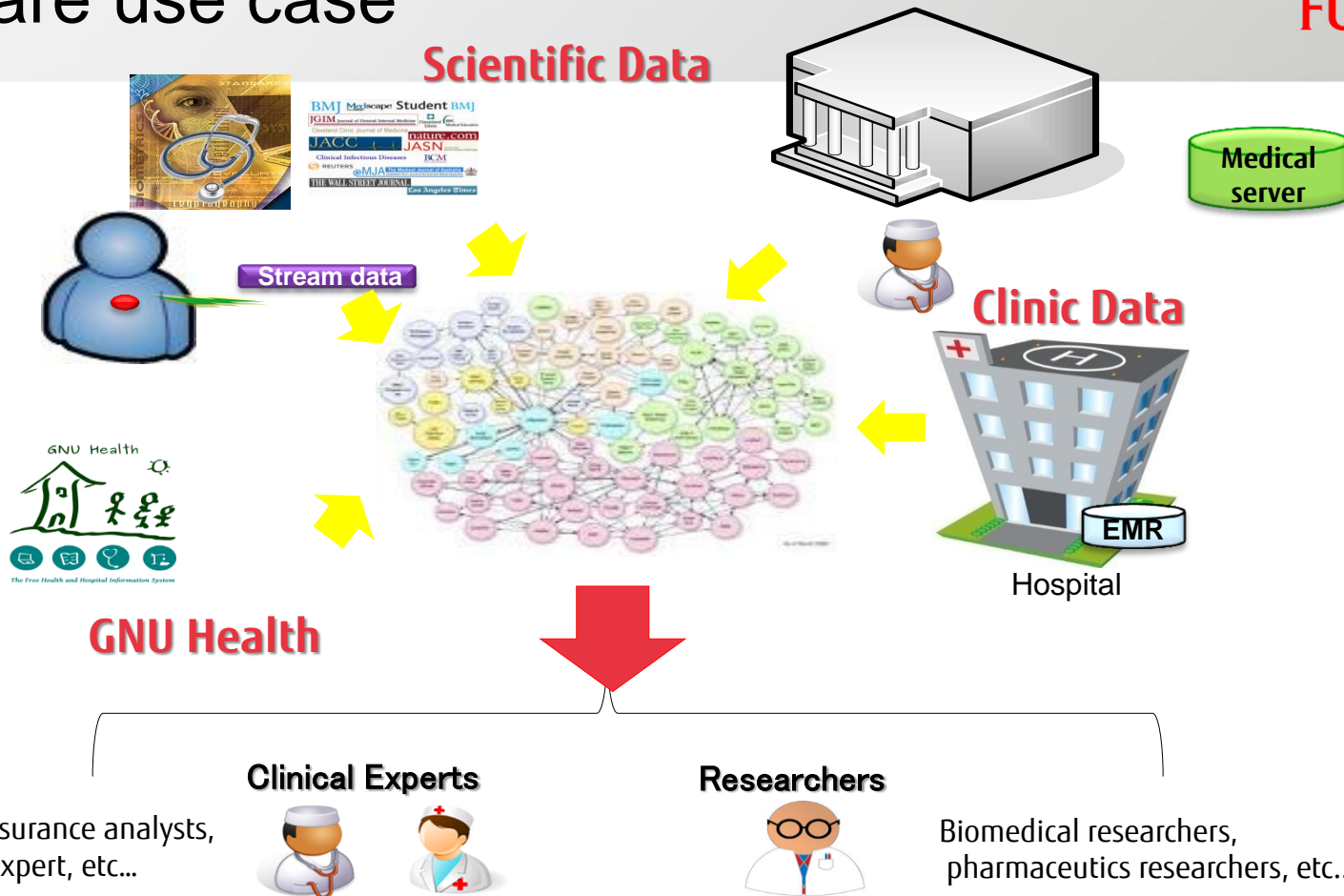
- Facilitation of serendipitous discovery when looking for known or possible effects of molecules tested as new drugs
- Easy identification of clinical trials relevant to a drug/protein/gene/disease scientists or clinicians are concerned with
- Provide new insights using symbolic analysis of the clinical trial data in connection to relevant scientific content

## ■ Technical features

- Technology for linking and aggregate stream and static data
- Technology for categorise particular knowledge patterns



# Healthcare use case

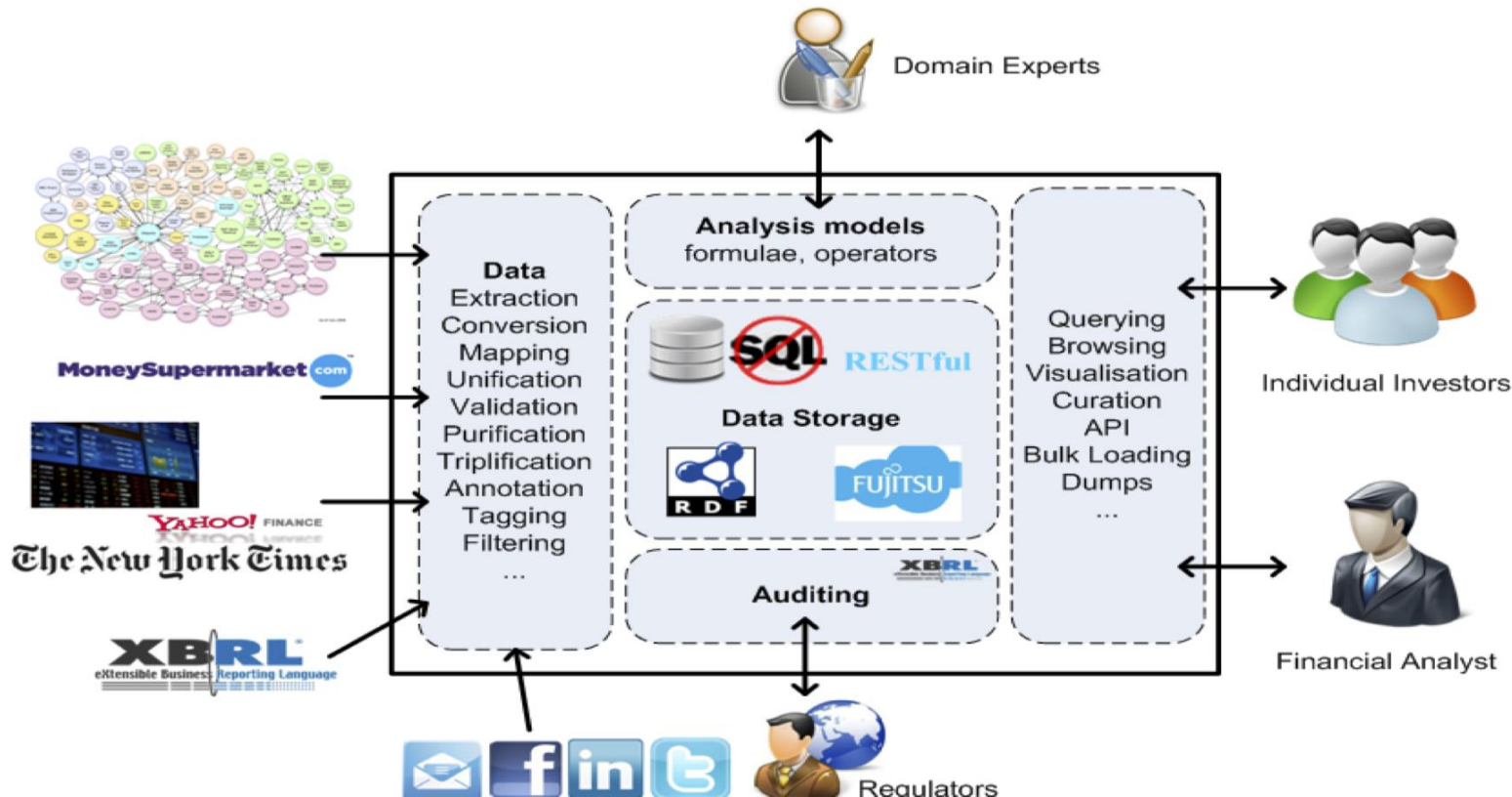


- Co-relate data from different sources and formats
- Avoid converting and importing some data
  - It's not desirable to convert to RDF data streams such as an ECG
  - But... each raw ECG stream must have an RDF representation for linking to other resources
- Timed queries and analytics
  - Analyse data within a specific time slot
- Run analytics on heterogeneous storage systems, e.g.
  - Analyse ECG on HBase
  - Co-relate Scientific and Clinic documents on Solr
  - Find relations between documents (authors, doctors, topics, ...) on Virtuoso

# FINANCIAL

- Purpose: **Company's Performance Comparison**
- Generate financial and market Linked Data
  - Mapping/Linking XBRL documents for generating Financial Key Performance Indicators (KPIs)
  - Linking Stock prices
  - Mapping taxonomies for company information with the identification code from LEI (Legal Identity Identifier)
  - Linking NewsML
- Effect of using Linked Data
  - Uniform management of different data sources by unique IDs
  - Investors can analyze various data for each purpose
- Technical features
  - Technology for linking **data of various standardizations**
  - Technology for linking **data updated frequently**

# Financial use case



- Co-relate data from different sources and formats
  - CrunchBase
  - DBPedia
  - NYTimes
  - ...
- Consume data streams from social media
  - LinkedIn
  - Twitter
- Timed queries and analytics
  - Analyse data within a specific time slot
- Run analytics on heterogeneous storage systems
  - All data on its native store, using RDF to annotate and link the data

# LIMITATIONS AND REQUIREMENTS


# Current challenges of use cases

- Data streaming
- Consuming data from heterogeneous data sources
- Storing data using heterogeneous technologies
- Analysing data using the available native storage mechanisms
  - While still using RDF as the main format for connecting all data (and representing most of it)



# Limitations and requirements

- Lack of systems that focus on the coordination of several big data technologies
  - How to benchmark an RDF based system that uses external technologies for specific purposes (e.g. full text search, map/reduce)
- Standardized benchmarking of automatic data conversion, integration and linkage through RDF
  - Data may need to be converted on-the-fly
  - **Current benchmarks do not take this into account, but...**
- Support for data streams
  - Bootstrap data generation + Stream data generation during the benchmark execution?
  - **Conversion, integration and linkage of data streams is part of our execution flow**



**FUJITSU**

shaping tomorrow with you