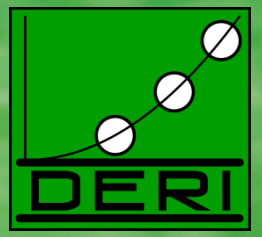


Challenges Ahead For Converging Financial Data

Edward Curry[1], Andreas Harth[2], Sean O'Riain[1]

[1] DERI, NUI Galway, Ireland
[2] Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB), Karlsruher Institut für Technologie (KIT)

W3C Workshop on Improving Access to Financial Data on the Web

October 2009, Arlington, Virginia USA

Agenda

- Motivation - Financial Data Ecosystem
  - Data Providers
  - Data Formats
  - Data Consumers
- Converging Financial Data from Multiple Sources
  - Entity Centric Approach
  - Architecture
  - Identity Mismatch
  - Data Query
- Data Integration Challenges
- Recommendations

Financial Data Ecosystem

Information Providers

Raw Data

?

Information Consumers

Financial Information Providers

- Individuals: e.g. CEOs reporting equity sale
- Companies: e.g. 10-K filing
- NGOs: e.g. sector-wide lobbying groups
- Government: e.g. regulators, central banks, statistics offices
- Worldwide organisations: UN, OECD
- Academics: various economists, public policy
- ...

- Publicly available datasets, purchased datasets or in-house sources

Various Data Formats

- Unstructed Text
  - News articles, press releases, raw transcripts of investor calls
- Hypertext
  - Coporate websites, goverment websites, ...
- Spreadsheets, et al.
  - CSV files, word docs, pdf, powerpoint, ...
- Strucutred Data
  - XML, XBRL, CSV, SDMX, ...
- Graph Structured Data in RDF
  - DBPedia, CrunchBase, RSS-CB, ...

Financial Information Consumers

- Competitive Analysis
  - Mash-up of financial figures and analyst commentary for decision support
- Regulatory Compliance
  - Forensic Economics
  - Spotting patterns or conditions that support fraud or money laundering
- Investment Analysis
  - Individual/Institutional investors
  - Transparent fund comparisons
  - Evaluate potential fund return

Goal

- Integrate data for:
  - Central access
  - Cross document analysis
- Our group works in data integration and have applied our approach to pilots in the financial services industry
- Report on experiences and lessons learned

Converging Financial Data from Multiple Sources

- Provide common data platform for search, browsing, analysis, and interactive visualisations across sources
- Entity centric approach
  - Single data view allowing information filtering and cross analysis
  - Consolidate data into coherent graph 'mashed up' from potentially thousands of sources
- Key challenge is semantic integration of structured and unstructured data from the open Web and internal corporate data sources

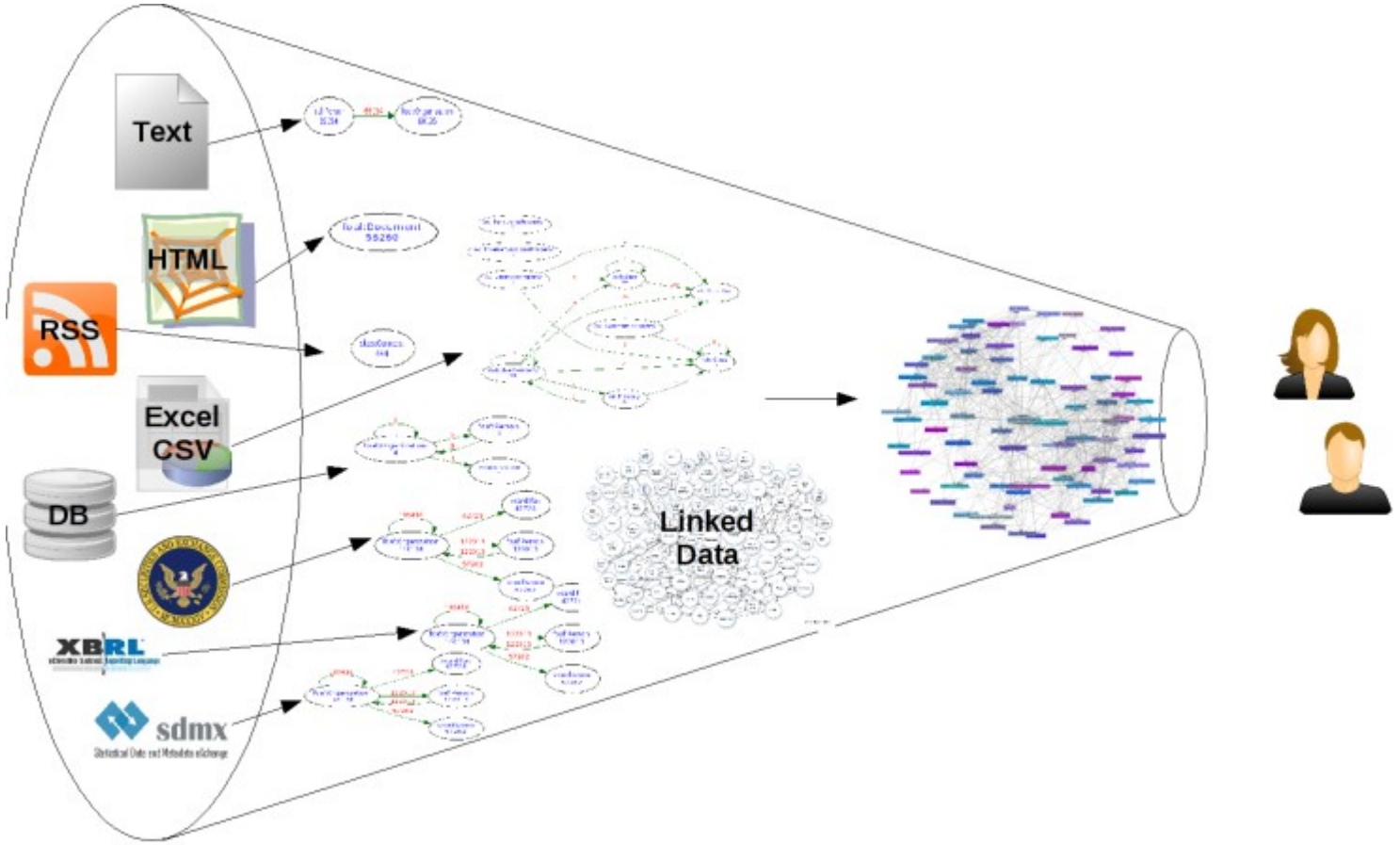
Converging Financial Data from Multiple Sources

- Large graph of RDF entities
- Entities typed according to what they describe
  - People, locations, organizations, publications as well as documents
  - Inter-relations and structured descriptions of entities
- Entities have specified relations to other entities
  - People can **work for** companies, people **know** other people, people **author** documents, organisations are **based in** locations, and so on

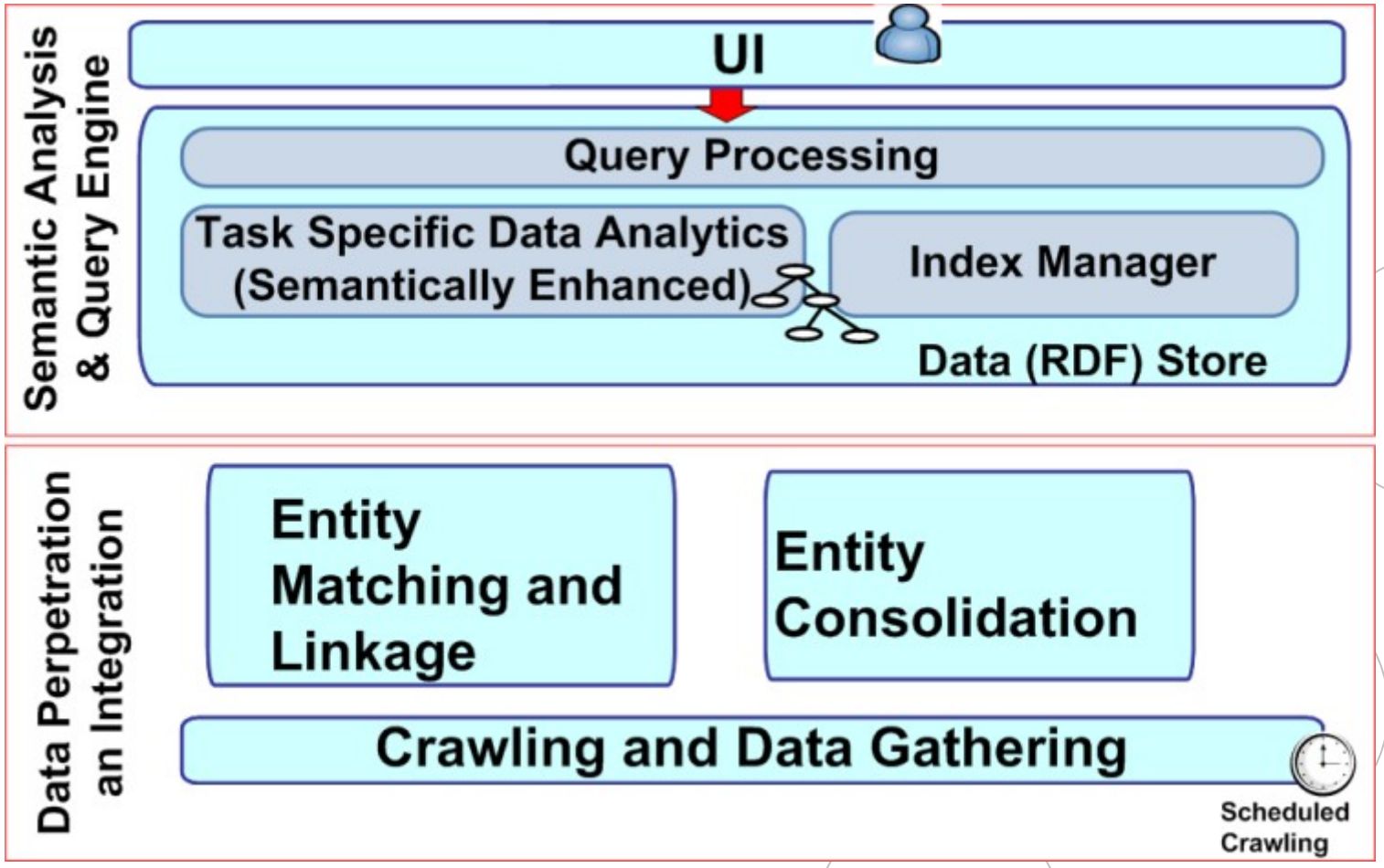
Data Integration Approach

- Lifting data sources to common format, in our case RDF (Resource Description Format)
- Integrating the disparate datasets into a holistic dataset by aligning entities and concepts
- Run domain/task specific analysis algorithms on integrated data
- Interactive browsing and exploration of integrated data or results of algorithmic analysis

Data Integration Approach



Architecture



Identity Mismatch

- Need to connect sources that may describe the same data on a particular entity

- Case studies analyzing connections between people and organizations
  - SEC filings (Form 4) identified 69K people connected to 80K organizations
  - Same analysis on database describing companies produced 122K people connected to 140K organizations
  - Data needed to be enrich and interlinked using entity consolidation (a.k.a. object consolidation) to avoid having the knowledge split over numerous instances
  - Ontology-based disambiguation

Data Query

- SPARQL, the semantic query language allows queries/questions to be asked:
  - What do the companies ‘Microsoft’ and ‘IBM’ have in common?
  - What competitors of ‘HP’ are in ‘Arlington’?
  - What’s the relationship between ‘Microsoft’ and ‘IBM’?

Data Integration Challenges

- Text/Data Mismatch
  - Human language often ambiguous
  - Same company might be referred to in several variations (e.g. IBM, International Business Machines, Big Blue)
  - Ambiguity makes cross-linking with structured data difficult
- Object Identity and Separate Schema
  - Sources differ in how they state the same fact
  - Differences on level of individual objects and schema
  - SEC Central Index Key (CIK) to identify people (CEOs, CFOs), companies, and financial instruments
  - DBpedia use URIs to identify same entities
  - Methods have to be in place for reconciling different representations of objects and schema

Data Integration Challenges

- Abstraction Levels (Data Context)
  - Financial data sources provide data at incompatible levels of abstraction
  - Classify data in taxonomies pertinent to a certain sector
  - Differences in legislation on book-keeping (e.g. Indicators from Euro regulators may not be directly comparable with indicators from US-based regulators)
  - Differences in geographic aggregation (e.g. region data from one source and country-level data from another, IBM Ireland Ltd, IBM Europe, IBM Global,...)

Data Integration Challenges

- Data Quality
  - General challenge integrating data from multiple sources
  - Errors in signage, amounts, labelling, and classification can seriously impede utility of systems operating on such data
  - Combining erroneous data aggravates the problem
  - Within open environment data aggregator has little or no influence on the data publisher
  - Challenge for data publishers/consumers to coordinate to fix problems in data or blacklist sites providing unreliable data

Recommendations

- Agree approach to the specification and use of common identifiers or at least their mappings
- Adhering to common publishing method reduces integration effort and facilitates data reuse
  - Linked Data principles
- Convergence between data providers requires coordination and time
  - No need for “Big Bang” integration
  - Follow a pay-as-you-go iterative approach to integration

Thank you for listening