# Ontological Integration of Information Extracted from Multiple Sources

Adam Funk, Diana Maynard, Horacio Saggion and Kalina Bontcheva
Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
Sheffield, S1 4DP, U.K.
{a.funk,d.maynard,h.saggion,k.bontcheva}@dcs.shef.ac.uk

## Abstract

We describe here an ontologically based approach to multi-source, multilingual information extraction. Structured, semi-structured and unstructured documents of various types are annotated using a range of hand-crafted and machine-learning information extraction processes; the resulting annotations are used as statements to update a knowledge base for business intelligence. Our approach in particular uses domain-oriented ontologies that extend the *de facto* Proton standard to ensure compatibility between the extracted data so that they can be integrated into a consistent, precise set of results.

## 1 Introduction

Multi-source information extraction typically deals with the specific problem of cross-document coreferencing, i.e. determining which named entities in a set of documents have the same referents; as Bagga and Baldwin [2] point out, this problem differs significantly from coreference identification within individual documents, where we can expect more consistency and a smaller potential domain. Approaches to this coreferencing problem include vector space modelling on document contexts [2, 3], adaptation of a Context Thesaurus originally developed for query refinement in information retrieval [18], and shallow syntactic analysis of multi-word terms [10].

Systems aimed at the business domain include JV-FASTUS [1], which carried out shallow text analysis with results that were interesting but (naturally for MUC) based on template-completion without reference to a domain-related ontology, and the MBOI tool [8] for discovering information about business opportunites on the internet, which however requires specific semi-structured data sources.

A semantically enhanced system is h-TechSight [14, 15], which uses information extraction and retrieval with an ontology to monitor markets and detect trends and changes, e.g. for business intelligence about competitors' products in company reports and news articles or for employers and applicants to watch the employment market. Unlike the system we will present here, however, the ontology is quite small with a few fixed concepts.

As Maynard et al. [16] point out, however, existing systems that aim to extract information for business intelligence do not deal sufficiently with unstructured text input. We therefore aim to develop and combine tools for various input types so that we produce coherent, consistent output.

## 2 Background

In the MUSING[1] project, we wish to provide a new generation of versatile yet integrated tools for business intelligence using semantically enhanced information extraction and reasoning for three application areas:

- financial risk management, especially credit risk management concerning small and medium enterprises (SMEs);
- internationalization, i.e. identifying, capturing, representing and localizing knowledge in the context of global competition; and
- operational risk measurement for IT systems.

We are interested in making the best use of declarative and statistical information extraction techniques on a variety of documents with different degrees and types of structure and mixtures of

[1] http://www.musing.eu/

numeric and textual content, such as companies' web pages, articles from the financial press, government documents and corporate financial reports. We have therefore designed a high-level approach to multi-source information extraction, based on integrating the results of various information extraction tasks using semantic knowledge.

# 3 Methodology

This section describes our ontologically-based approach to the problem of integrating information extraction from diverse sources using various information extraction techniques.

## 3.1 Input

We wish to extract information from a variety of document types which present different problems and characteristics for information extraction.

News articles consist mainly of free natural language text, with some metadata from the provider's database as well as XML or HTML annotation. Companies' web pages (particularly the index, "contact us" and "about us" pages of each site examined) similarly consist of free text with varying degrees of HTML annotation, some of which (such as headings and URLs) can be particularly useful for information extraction.

Wikipedia[2] articles are also mostly free text, although parallel articles often have parallel structure and tabular data in regular formats (for example, each article about a country or region usually contains a fairly standardised table with figures for population, surface area, etc., and similar headings and natural-language expressions recur). The CIA World Factbook[3] has a much more consistent and therefore easy analysable format (but does not cover many regions within countries). Government documents also contain a wide variety of numeric and textual information in semi-structured and unstructured forms.

Balance sheets and other financial reports are now structured fairly consistently according to international accounting standards and can also be written in the emerging XBRL[4]. [9] However it is also useful to take advantage of NLP techniques to analyse the information in the free-text notes to these reports, which may significantly affect the interpretation of the easily analysable numeric parts.

It is also worth noting that to meet the needs of modern business intelligence we wish to take advantage of sources in various languages.

## 3.2 Extraction techniques

Our information extraction applications are based primarily on GATE[5], which provides a development environment, an architecture, a library of robust, adaptable tools for natural language processing (including machine-learning), and facilities for manual annotation of documents, and which is well-suited for multilingual information extraction. [4, 6]

These applications fall into two categories: *declarative* and *machine-learning.*

### 3.2.1 Declarative tools

Our declarative or hand-crafted applications are generally derived from GATE's standard information extraction system, ANNIE [13], which provides standard NLP tools (tokeniser, sentence splitter, POS (part-of-speech) tagger, lemmatiser) as well as some gazetteers and JAPE[6] grammars for general-purpose information extraction. ANNIE already has very good performance (F-measure 92.9%) for traditional information extraction on general news texts [16] and is therefore a good base to build on.

For this purpose, we add gazetteers of key words and phrases found in the documents and JAPE grammars to detect patterns and annotate the information desired by our representative users in the project. Table 1 lists a small sample of the datatypes requested for commercially evaluating different countries and regions. We are developing several applications along these lines to deal with the various input document types.

Tabular data can be analysed with gazetteers and JAPE grammars designed to identify the row and column headings and boundaries and to annotate the statistics accordingly. JAPE rules can also take advantage of a document's "original markups" such as HTML or XML tags, and therefore treat headings differently from paragraphs, for example.

### 3.2.2 Machine learning

Users are also manually annotating documents using GATE's Ontology-based Corpus Annotation

[2] http://www.wikipedia.org/

[3] https://www.cia.gov/library/publications/the-world-factbook/index.html

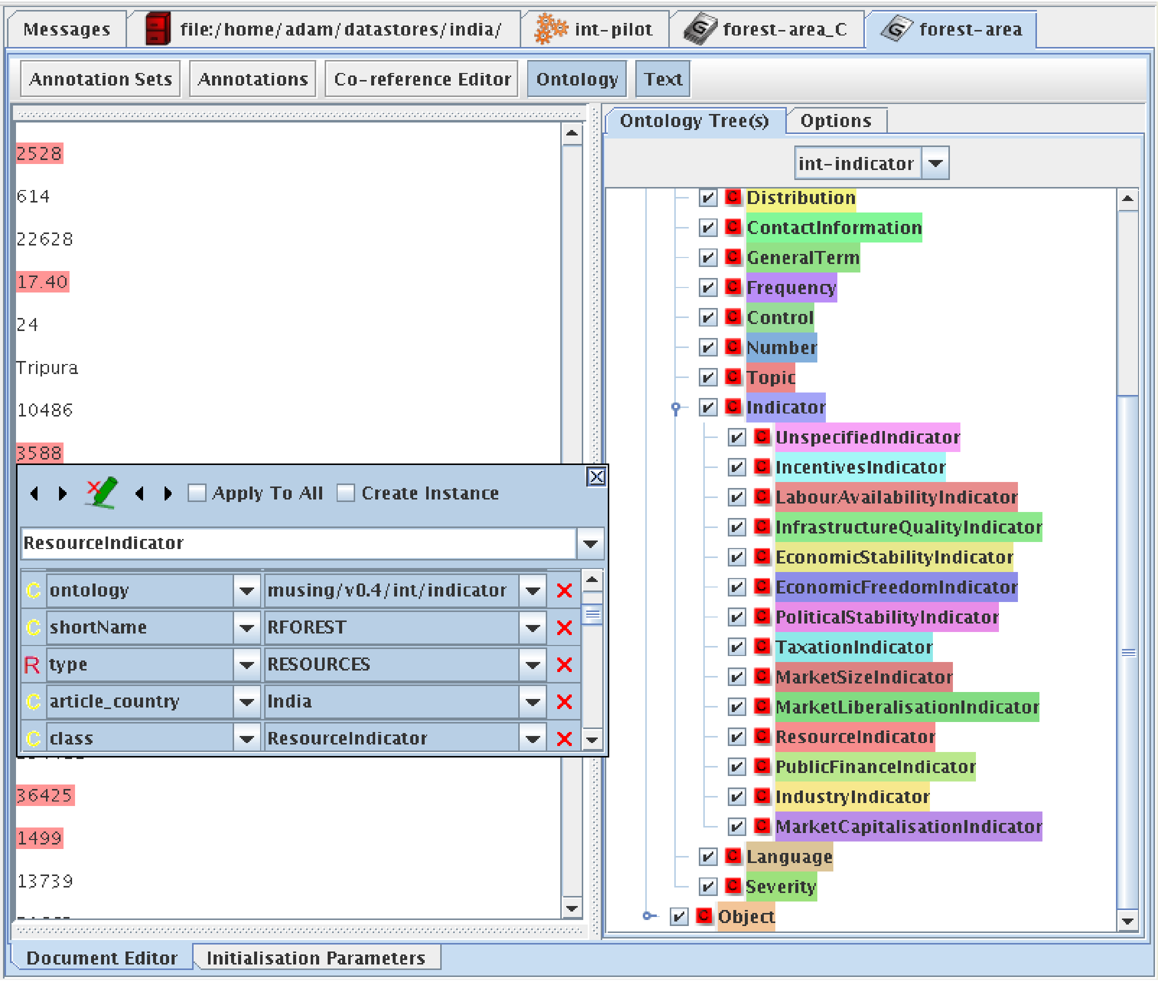
[4] eXtensible Business Reporting Language http://www.xbrl.org/

[5] http://gate.ac.uk/

[6] JAPE is an engine for pattern-matching over annotations on GATE documents and adding further annotations or executing arbitrary Java code. [5]

| Class | Short name | Full name |
|---|---|---|
| LabourAvailabilityIndicator | EMP | Employment rate |
| | WAGE | Minimum wage |
| MarketSizeIndicator | RUR | Rural population (%) |
| | LRT | Literacy rate total (%) |
| | DENS | Population density |
| ResourceIndicator | FOREST | Forest area (%) |
| | RFOREST | Reserved forest (sq km) |
| | AGRIC-LAND | Agricultural land (%) |

**Table 1:** *Example datatypes for region evaluation*



2528

614

22628

17.40

24

Tripura

10486

3588

36425

1499

13739

ResourceIndicator

| | | |
|---|---|---|
| C | ontology | musing/v0.4/int/indicator |
| C | shortName | RFOREST |
| R | type | RESOURCES |
| C | article_country | India |
| C | class | ResourceIndicator |

**Figure 1:** *Ontological annotation in GATE*

Tool, OCAT, as illustrated in Figure 1.[7] Although GATE's annotation model (which is based on the TIPSTER model) allows each annotation to contain a map of arbitrary feature-value pairs, using the OCAT extension constrains all new manual annotations to have the same type (usually `Mention`) and a `class` feature whos values must be selected from the active ontology. The annotators can do this work using the normal, locally installed GATE GUI or an easy remote service based on JavaWebStart[8].

These documents (at this point, especially the news articles from the financial press) are being used to train machine-learning applications for the project's information extraction tasks.

In particular, the web service allows us to serve documents that have already been automatically annotated so that the human annotators can correct them by adding, removing or changing annotations. These documents with manually improved annotations can be fed back into an information extraction system designed for progressive improvements in machine learning that takes advantage of the ontological structure of the annotations. [11, 12]

#### 3.2.3 XBRL mapping

Financial information written in XBRL (which was briefly introduced in Section 3.1) is already machine-readable, and MUSING's ontologists are working on mappings between XBRL and our domain ontologies to ensure that such data can be annotated very precisely with full integrability into our system. Related information supplied with XBRL data (such as free text associated with a financial statement) can be annotated with the other techniques (declarative information extraction and machine learning) using the same ontologies for consistency (although at a performance level appropriate for analysis of natural language). [7]

### 3.3 Integration

Instead of concentrating on the traditional, low-level multi-source information extraction tasks such as cross-document coreferencing, we are interested here in the high-level task of refining and growing a knowledge base in a consistent manner.

For this purpose, ontology experts at DERI Innsbruck[9] have developed and continue to refine, based on information provided by other MUSING partners, a set of domain ontologies for business intelligence that extend the Proton[10] ontology.

To be precise, the MUSING ontologies contain `owl:imports` statements that refer to Proton's *System*, *Top* and *Knowledge Management* modules (using the *Upper* module would adversely affect decidability) so that our domain-specific extensions consist of subclasses and instances of Proton classes, as well as instances of our classes; this extension of a well-known *de facto* standard in the semantic web field could facilitate interoperability with other parties' tools in the long term.

We therefore ensure that all the automatic annotation carried out by both types of components described in Section 3.2 makes good use of this semantic enhancement: specifically, we design our declarative components so that every annotation contains `ontology` and `class` features whose values point to a particular MUSING ontology and to one of its classes, respectively; and we ensure that our training data (manually annotated documents) and machine-learning components also respect this requirement.

We can process a diverse range of input documents through appropriate information extraction engines in a many-to-many relationship; documents can be analysed with several techniques to capture a wider range of information.

Figure 2 shows a schematic of the integration of these processes into a multi-source information-extraction *system*, in which the common set of domain ontologies (tied together as extensions of Proton) serves to unify the engines (as indicated by the dashed arrows) so that the resulting annotated documents are semantically coherent and the information can be consistently added to an ontological knowledge base.

The different applications running in parallel can also be evaluated to compare their performance as part of the evaluation of our system and to help us refine it.

In effect, the constraints on annotation according to the domain ontologies act as an "information funnel" to ensure consistency and compatibility of the extracted information going into the knowledge base.

### 3.4 Coreferencing

Our approach also deals with the classical problem of cross-document coreferencing, but takes advantage of the semantically enhanced annotation in order to treat it as an *ontology population* problem. For example, Figure 3 shows three texts

[7] The right-hand pane shows ontology classes colour-coded to match the corresponding annotations in the left-hand pane.

[8] `http://java.sun.com/products/javawebstart/`

[9] `http://www.deri.at/`

[10] `http://proton.semanticweb.org`

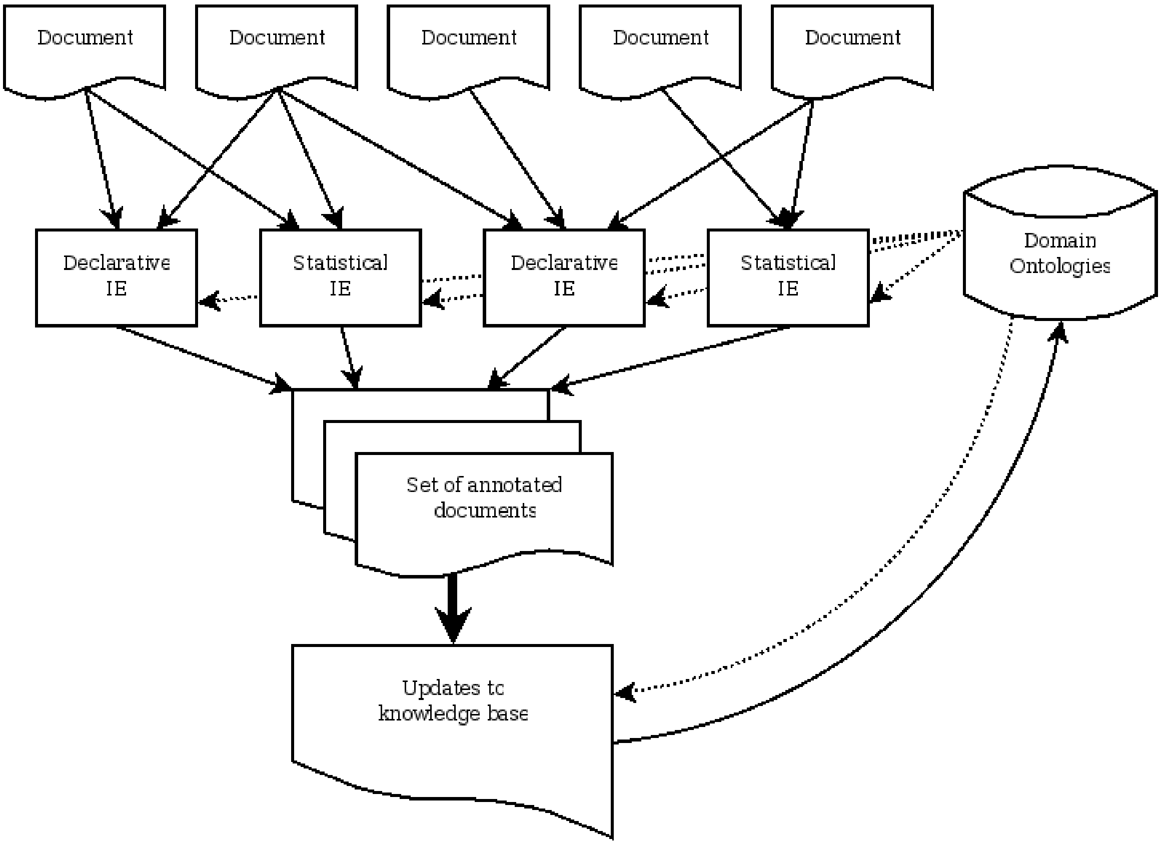


**Figure 2:** *Ontological integration of extracted information*

that use different expressions to refer to the same company (Alcoa Inc.), and we wish to link the separate and often complementary pieces of information (address, chairman, financial announcements) together for better intelligence.

We treat each named entity as a possible ontology instance and retrieve candidate instances (of companies, for example) from the relevant domain ontology, along with known features of those instances from the knowledge base. We will then employ a rule-based system defined by domain experts to compute similarity scores between the possible instance and the candidates, in order either to dereference the named entity or to add a new instance to the ontology (which uses the KIM OWLIM [17] semantic repository).

### 3.5 Refinement

Both the declarative and machine-learning information extraction tools will be continually refined through a feedback loop in which human annotators correct the automatic annotations on selected documents using AnnotatorGUI [reference], a web application tool recently developed by the NeOn project.[11]

This application is deployed as a JavaWebStart service (briefly discussed above in Section 3.2.2) which runs and loads GATE documents and ontologies from a server at the University of Sheffield and save the modified documents back to the same server, where they can later be manually inspected to refine the declarative information extraction tools or automatically fed back into a machine-learning loop. The correct annotations (according to the human annotator) and the previous automatic annotations can be stored in distinct annotation sets in the same document, so that the automatic ones can be scored using GATE's AnnotationDiff tool.

New documents can also be introduced into the loop in order to test and improve the versatility of existing IE tools and to determine if new ones are required to enlarge the scope of the integrated system.

## 4 Discussion and future work

We have described the design of a coherent system of information extraction for business intelligence, which uses expertly designed domain ontologies that extend a *de facto* standard (Proton) for the semantic web in order to integrate the output of a variety of separate information extraction tools that process a variety of document types ranging from unstructured text to highly structured information. This system also allows superficial redundancy in that each document could be processed using multiple tools in order to improve the overall precision (i.e. to reduce the quantity of "missed" information).

We have implemented much of this system but have not yet annotated enough data to carry out reliable quantitative evaluation[12], which will be a focus for our work in the new future—not just to validate our work but also to continue to carry it out, since such evaluation is an important part of the feedback loop described in Section 3.5.

## Acknowledgements

This research is supported by the EU grant IST-2004-027097 for the MUSING (Multi-industry, semantic-based next generation business intelligence) project and benefits from work supported by EU grant IST-2005-027595 for the NeOn (Networked ontologies) project.

## References

[1] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, and M. Tyson. Description of the JV-FASTUS system as used for MUC-5. In *Proceedings of the Fourth Message Understanding Conference MUC-5*, pages 221–235. Morgan Kaufmann, California, 1993.

[2] A. Bagga and B. Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85, 1998.

[3] A. Bagga and A. W. Biermann. A methodology for cross-document coreference. In *Proceedings of the Fifth Joint Conference on Information Sciences (JCIS 2000)*, pages 207–210, 2000.

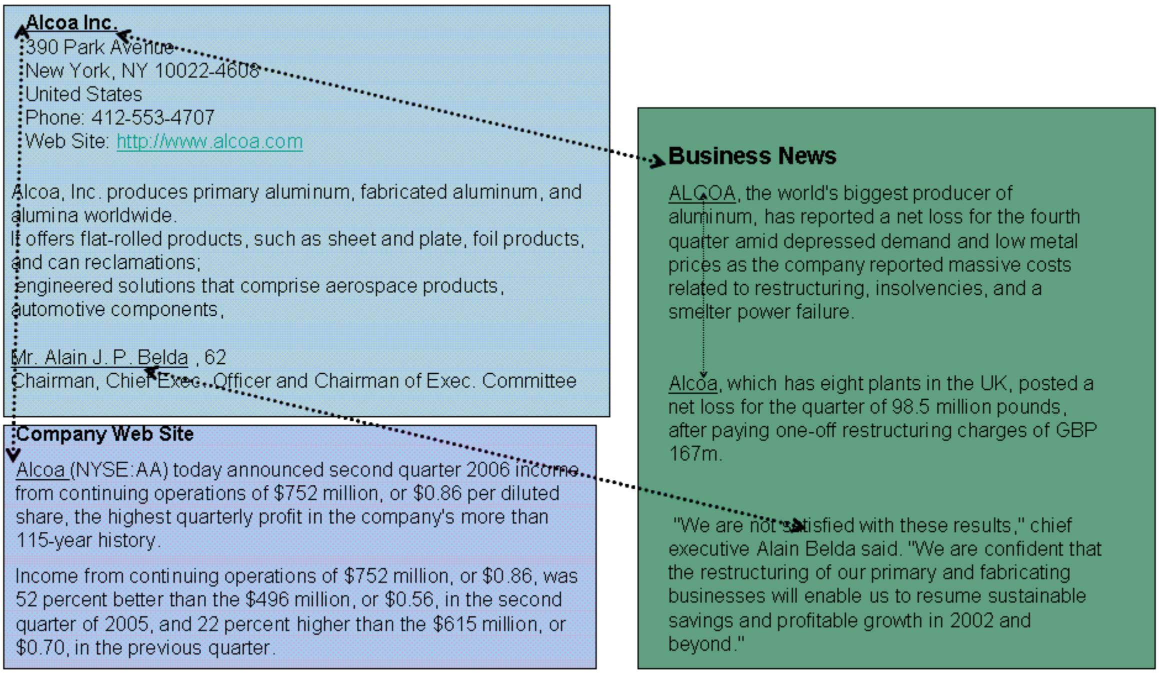
[4] K. Bontcheva, D.Maynard, V. Tablan, and H. Cunningham. GATE: A Unicode-based infrastructure supporting multilingual information extraction. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria, 2003.

[5] H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, Nov. 2000.

[6] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[11] http://www.neon-project.org/

[12] This will be based on precision and recall over annotations, as is usual in this type of work, with appropriate scoring for partial matches not only for overlapping annotation wspans but also for subsumption of annotation classes.



**Alcoa Inc.**
390 Park Avenue
New York, NY 10022-4608
United States
Phone: 412-553-4707
Web Site: http://www.alcoa.com

Alcoa, Inc. produces primary aluminum, fabricated aluminum, and alumina worldwide.
It offers flat-rolled products, such as sheet and plate, foil products, and can reclamations;
engineered solutions that comprise aerospace products, automotive components,

Mr. Alain J. P. Belda , 62
Chairman, Chief Exec. Officer and Chairman of Exec. Committee

**Company Web Site**

Alcoa (NYSE:AA) today announced second quarter 2006 income from continuing operations of $752 million, or $0.86 per diluted share, the highest quarterly profit in the company's more than 115-year history.

Income from continuing operations of $752 million, or $0.86, was 52 percent better than the $496 million, or $0.56, in the second quarter of 2005, and 22 percent higher than the $615 million, or $0.70, in the previous quarter.

**Business News**

ALCOA, the world's biggest producer of aluminum, has reported a net loss for the fourth quarter amid depressed demand and low metal prices as the company reported massive costs related to restructuring, insolvencies, and a smelter power failure.

Alcoa, which has eight plants in the UK, posted a net loss for the quarter of 98.5 million pounds, after paying one-off restructuring charges of GBP 167m.

"We are not satisfied with these results," chief executive Alain Belda said. "We are confident that the restructuring of our primary and fabricating businesses will enable us to resume sustainable savings and profitable growth in 2002 and beyond."

**Figure 3:** *Related information from multiple sources*

[7] T. Declerck and H. Krieger. Translating XBRL into Description Logic: an approach using Protege, Sesame and OWL. In *Proceedings of Business Information Systems (BIS)*, Klagenfurt, Germany, 2006.

[8] J.-Y. N. F. Paradis and A. Tajarobi. Discovery of business opportunities on the internet with information extraction. In *Workshop on Multi-Agent Information Retrieval and Recommender Systems (IJCAI)*, pages 47–54, Edinburgh, Scotland, 2005.

[9] F. Fornasari, A. Tommasi, C. Zavattari, R. Gagliardi, T. Declerck, and M. Nannipieri. Xbrl web-based business intelligence services. In P. Cunningham and M. Cunningham, editors, *Innovation and the Knowledge Economy: Issues, Applications, Case Studies. Proceedings of eChallenge 2005*. IOS Press, 2005.

[10] Z. Kazi and Y. Ravin. Who's who? Identifying concepts and entities across multiple documents. In *Proceedings of the* 33$^{rd}$ *Hawaii International Conference on System Sciences*, volume 3, Hawaii, USA, Jan 2000.

[11] Y. Li, K. Bontcheva, and H. Cunningham. Perceptron-like learning for ontology based information extraction. Technical report, University of Sheffield, Sheffield, UK, 2006.

[12] Y. Li, K. Bontcheva, and H. Cunningham. Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction. In *16th International World Wide Web Conference (WWW2007)*, pages 777–786, 2007.

[13] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. Named Entity Recognition from Diverse Text Types. In *Recent Advances in Natural Language Processing 2001 Conference*, pages 257–274, Tzigov Chark, Bulgaria, 2001.

[14] D. Maynard, H. Cunningham, A. Kourakis, and A. Kokossis. Ontology-Based Information Extraction in hTechSight. In *First European Semantic Web Symposium ( ESWS 2004)*, Heraklion, Crete, 2004.

[15] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Apects of the Semantic Web")*, Heraklion, Crete, 2005.

[16] D. Maynard, H. Saggion, M. Yankova, K. Bontcheva, and W. Peters. Natural Language Technology for Information Integration in Business Intelligence. In *10th International Conference on Business Information Systems (BIS-07)*, Poznan, Poland, 2007.

[17] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM – Semantic Annotation Platform. *Natural Language Engineering*, 2004.

[18] Y. Ravin and Z. Kazi. Is Hillary Rodham Clinton the president? disambiguating names across documents. In *Proceedings of the ACL 1999 Workshop on Coreference and its Applications*, Jun 1999.