

Finding ECS Alumni, Semantically

Matthew Wilson

July 28, 2005

1 Introduction

The AKT project was approached for help locating alumni from ECS. It is estimated that there are approximately three thousand graduates from the School since its inception as the Department of Electronics in 1947[2]. The School has no active contact with most of these graduates. However, it is known that some of these people hold prestigious positions in industry and academia. If they could be located, they could be a great help to marketing in the School.

It was hoped that a semantic web application could be developed to help locate these alumni, leveraging some of the technologies researched and developed by the AKT IRC. Data collected could then be reused for other AKT activities and to inform other research.

Initially, a feasibility study was conducted to determine whether semantic web technologies could help solve this problem. At this stage, it was thought that no data was available on ECS alumni whatsoever. Therefore, the proposed approach consisted of using a search engine, attempting to locate individuals with a degree from ECS. An example query is given below. Such queries did return some relevant results, but also a large number which were irrelevant. These results could be readily identified by a human agent, but discriminating programmatically is not trivial. Using natural language processing techniques and information extraction tools such as GATE[3] was investigated. This would require the development of a system such as used in the Artequakt project[4] which is capable of ontology-based entity relation extraction in order to identify people who gained degree from ECS, and the institutions or companies for which they now work. However, it was decided that this was beyond the scope of this project.

Example Query:

```
( studied OR degree OR graduated) ("university of  
Southampton" OR "Southampton University") (ECS OR "electronic  
engineering" OR "electrical engineering" OR "computer  
science" OR "software engineering")
```

It was later discovered that a large amount of data is published by the University Alumni Office for those alumni who are lost those who the Alumni Office is not in

contact with. Their website¹ contains the names, years of graduation and departments for roughly 25,000 graduates in all disciplines. With this data as a starting point, the application would only need to attempt to locate their current whereabouts and position. Again, the approach of using a search engine with these names was examined, but similar problems were encountered to those explained above. Therefore, a similar approach was adopted to data collection as for CS AKTive Space[5]: locate and map data from heterogeneous structured data sources into a common ontology. Using this approach, locating alumni is merely a matter of identifying coreferent entities between the various datasets.

At this point it was clear that a solution was viable and would be developed in the following stages:

1. Develop an ontology to represent alumni
2. Extract data from various sources and map it to a common ontology
3. Develop a front-end application to allow a user to identify alumni

2 Ontology Development and Data Storage

The majority of the information about alumni can be expressed in terms of the properties defined for people by the AKT reference ontology². However, a few additional properties were required to express the former name, year of graduation and department. The URIs for each alumnus were created by concatenating the first name, last name and year of graduation as this has a reasonable probability of being unique.

The natural choice for data storage was 3store[1], the RDF triplestore developed within AKT. 3store provides efficient bulk storage on a MySQL back end, and is highly scalable. It provides a command line tool for RDQL queries, and both OKBC and RDQL over HTTP are supported. Additionally, a Perl module is provided for interfacing with the command line tool.

3 Data Harvesting

The first task was to gather data from the Alumni Office website, express it in terms of the ontology and assert it in 3store. The data is made available in a separate file for each year of graduation. Some of these files are in PDF format, some are Microsoft Word documents and some HTML pages. It was decided that as there are relatively few PDF files, they would be ignored as they are very difficult to parse. The parser for the Word Documents and HTML pages was written in Perl. The Word parser uses OLE to connect to Word, open the document and extract the relevant fields from the table. The HTML parser is a state machine using a Perl HTML token-parsing module. In both cases the parser produces an output file in N-Triples format. Rapper, an open

¹<http://www.soton.ac.uk/Alumni/LostAlumni>

²<http://www.aktors.org/ontology/portal>

source RDF parser distributed as part of the Raptor toolkit³ was then used to convert the data to RDF/XML to import into 3store.

The next task was to locate some structured data sources with which the alumni data could be correlating. This was not particularly easy. In the academic domain, a lot of work had previously been done in gathering data from RAE 2001. This data includes details of every UK academic and the institutions that they work for. This data is available in RDF format⁴ expressed in the AKT ontology and therefore seemed a logical choice. It also provides details of their research interest in terms of Units of Assessment. Mappings were created by hand from the Units of Assessment to the department fields from the alumni data. This also overcame the problem of the inconsistency of the department field. So RAE Unit of Assessment 25 (Computer Science) is mapped to "Computer Science" and "Electronics and Computer Science" in the alumni department field. This enables the user to select one subject (by RAE Unit of Assessment) and get all relevant results. However, there are some problems. The data is clearly somewhat out of date. It also does not provide a huge amount of information, in particular academics only have family names and initials.

As ECS are looking for alumni in positions of influence, it is also important to look for alumni in the industrial domain. Finding structured sources of data available in the public domain was difficult. The final decision was to use the Applegate directory⁵ which publishes details of thousands of top UK companies and their senior members of staff. This a very rich and useful dataset and will likely be usable future activities. A similar technique was used to gather data from the Applegate site as for the Lost Alumni site.

4 User Interface

It was decided that the front-end to the application should be web-based. Perl was chosen as the implementation language due to the ease of rapid development, good CGI support and the supplied module to interface with 3store. Due to the complexity of the queries involved, it was decided that queries against the RAE and Applegate data sets should be mutually exclusive. In testing, the queries were found to be too complex to be executed in a short enough time to return the results as a web page. So an asynchronous solution was chosen in which the CGI script spawns a daemon process on the server which will e-mail the results to the user when they are available.

5 Results and Performance

It is clear that there are a fairly large number of matches. However, identification of coreferent entities is only based on first name and last name for Applegate data, or first initial and last name for RAE data. This means it is far from certain that the two entities are the same person. In fact, it may be a possible for a human agent to easily

³<http://librdf.org/raptor/>

⁴<http://www.hyphen.info/rae.php>

⁵<http://www.thermoforming-specialist.co.uk>

Figure 1: Screenshot of web-based user interface

identify matches by eye which are implausible. For example, it is clear that someone graduating in 2004 is unlikely to be a managing director by 2005. However, this is not a discernment that can easily be made programmatically.

Another interesting matter is the time taken for the queries to execute. In both the case of the RAE query (Figure 3) and the Applegate query (Figure 2), a large Cartesian product between the two datasets must be calculated, which is a relatively expensive operation. This is exacerbated in the case of the RAE data as matching is done on first initial, and there may be many thousands of results for any one initial. Also, many properties are selected for each interesting URI. 3store cannot always predict whether a variable will be bound to a URI or a literal in these cases. This is characterised by a `LEFT JOIN` in the SQL, which is very expensive. Therefore it was necessary to split the operation into the main query (Figure 3) to return the interesting URIs (which takes approximately twenty minutes to execute), and subsequent queries (Figure 4) which return the required properties.

6 Conclusions

This paper has discussed the research, design and implementation of an application to investigate university alumni by correlating their details with publicly available data. This produced a reasonable number of plausible results. However, to decide whether they are actually coreferent entities is beyond the capabilities of the application and re-

```

SELECT ?firstname, ?lastname, ?year, ?subject, ?comp, ?job
WHERE (?x, <akt:given-name>, ?firstname)
      (?x, <akt:family-name>, ?lastname)
      (?x, <alumni:graduated-year>, ?year)
      (?x, <alumni:graduated-from-department>, ?subject)
      (?subject, <alumni:maps-to-rae-uoA>, <hesa:UoA-25>)
      (?p, <akt:family-name>, ?lastname)
      (?p, <Akt:works-for>, ?c)
      (?c, <akt:has-pretty-name>, ?comp)
      (?p, <akt:given-name>, ?firstname)
      (?p, <akt:has-job-title>, ?job)
USING alumni FOR <http://www.soton.ac.uk/Alumni/LostAlumni#>,
      akt FOR <http://www.aktors.org/ontology/portal#>
      hesa FOR <http://www.hesa.ac.uk/#>

```

Figure 2: Example query to correlate alumni with the Applegate Directory

quires the deductive reasoning of a human agent. It may be possible to conduct further verification by consulting the institution's or company's website for which an individual works and gathering further data. Due to the large number of possible formats of individual sites, some advanced NLP would be required. The application could be further extended if additional structured data sources were added to broaden the data sets to be correlated with. In particular, the application currently only has data pertaining to UK companies and academics, although it is reasonable to assume that some proportion of ECS alumni will now be working overseas. In general, the application is a success as it provides the ECS marketing team with position details for some of the alumni who may now be in positions of influence.

References

- [1] Harris, Stephen and Gibbins, Dr Nicholas *3store: Efficient Bulk RDF Storage* in Proceedings 1st International Workshop on Practical and Scalable Semantic Web Systems, Sanibel Island, Florida, USA.
- [2] University of Southampton, 2005 <http://www.ecs.soton.ac.uk/about/history/>
- [3] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002
- [4] Alani, Harith and Kim, Sanghee and Millard, David and Weal, Mark and Hall, Wendy and Lewis, Paul and Shadbolt, Nigel (2003) Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems 18(1):pp. 14-21

```

SELECT ?moreinfo, ?u
WHERE  (?moreinfo, <alumni:graduated-from-department>, ?z),
      (?z, <alumni:maps-to-rae-uoA>, <hesa:UoA-25>),
      (?moreinfo, <akt:family-name>, ?name),
      (?y, <akt:family-name>, ?name),
      (?moreinfo, <alumni:initial>, ?ali),
      (?y, <alumni:initial>, ?ali),
      (?y, <akt:has-research-interest>, <hesa:UoA-25>),
      (?y, <akt:works-for>, ?c)
      (?c, <akt:has-pretty-name>, ?u)
USING alumni FOR <http://www.soton.ac.uk/Alumni/LostAlumni#>,
      akt FOR <http://www.aktors.org/ontology/portal#>,
      hesa FOR <http://www.hesa.ac.uk/#>

```

Figure 3: Example of main query to correlate alumni with the RAE Data

```

SELECT ?year, ?formername, ?firstname, ?lastname, ?subject
WHERE  (?moreinfo, <akt:given-name>, ?firstname)
      (?moreinfo, <akt:family-name>, ?lastname)
      (?moreinfo, <alumni:former-name>, ?formername)
      (?moreinfo, <alumni:graduated-year>, ?year)
      (?moreinfo, <alumni:graduated-from-department>, ?subject)
USING akt FOR <http://www.aktors.org/ontology/portal#>
      alumni FOR <http://www.soton.ac.uk/Alumni/LostAlumni#>

```

Figure 4: Example of subsequent queries to retrieve extra properties of interesting entities find by the initial query on RAE data (Figure 3)

- [5] Glaser, Hugh and Alani, Harith and Carr, Les and Chapman, Sam and Ciravegna, Fabio and Dingli, Alexiei and Gibbins, Nicholas and Harris, Stephen and schraefel, m.c. and Shadbolt, Nigel (2004) CS AKTiveSpace: Building a Semantic Web Application, in Bussler, Christoph and Davies, John and Fensel, Dieter and Studer, Rudi, Eds. The Semantic Web: Research and Applications (First European Web Symposium, ESWS 2004), pages pp. 417-432. Springer Verlag.

Last Name	First Name	Year	Institution
Cowling	Paul	1981	University of Nottingham
Higgins	Peter	1981	University of Essex
McAllister	Hilton	1982	University of Ulster
Mitchell	Ian	1982	Middlesex University
Moore	Stephen	1982	University of Cambridge
Watson	David	1982	University of Sussex
Williams	Stephen	1982	University of Reading
Tan	Chin	1983	The Queen's University of Belfast
Campbell	John	1984	University College London
Campbell	John	1984	The Queen's University of Belfast
Davies	Nigel	1985	Lancaster University
Willis	Patrick	1985	University of Bath
Gray	Philip	1986	University of Aberdeen
Miller	Paul	1986	The Queen's University of Belfast
Parsons	Paul	1986	University of Hull
Williams	David	1986	Liverpool John Moores University
Jones	Richard	1987	University of Kent at Canterbury
Lee	Michel	1987	University of Birmingham
Lee	Michel	1987	University of Wales, Aberystwyth
Phillips	Colin	1987	University of Newcastle
Turner	Douglas	1987	University of Kent at Canterbury
Turner	Douglas	1987	Queen Mary, University of London

Table 1: Results from running the query (Figure 2) against the RAE data

Last Name	Year	Job Title	Company	First Name
Bassett	1980	Managing Director	Korn/Ferry International	Peter
Hart	1980	Director	Industrial Power Cooling (I P C) Ltd	Nicholas
Hunt	1980	Managing Director	Adec Propulsion Ltd	Jeremy
Hunt	1980	Director	Hunt Brothers	Jeremy
Spencer	1980	Commercial Director	Rowe Hankins Components Ltd	Peter
Spencer	1980	Managing Director	Atlantis International Ltd	Peter
Spencer	1980	Managing Director	Scotsman Beverage Systems	Peter
Webb	1980	Director	Intermusic Ltd	Richard
Webb	1980	Sales and Marketing Director	Hiatt Hardware Ltd	Richard
Brown	1981	Business Development Manager	Hi-Tech Mouldings Ltd	Richard
Brown	1981	Technical Manager	PNR Design Services	Richard
Brown	1981	Marketing Manager	Killgerm Chemicals Ltd	Richard
Brown	1981	Sales and Marketing Director	Hi-Tech Design Ltd	Richard
Brown	1981	Managing Director	South England Pastries	Richard
Brown	1981	Managing Director	Lwd Precision Engineering Co Ltd	Richard
Brown	1981	Managing Director	Encase Northern Ltd	Richard
Brown	1981	Managing Director	Decorative Glass By SGO	Richard

Table 2: Results from running the query (Figure 2) against the Applegate data