Revealing Trends and Insights in Online Hiring Market Using Linking Open Data Cloud: Active Hiring a Use Case Study

Amar-Djalil Mezaour[1], Julien Law-To[1], Robert Isele[3],
Thomas Schandl[2], and Gerd Zechmeister[2]

[1] EXALEAD, Dassault Systèmes: http://labs.exalead.com/
[2] Semantic Web Company GmbH: http://www.semantic-web.at/
[3] Freie Universität Berlin: http://www.fu-berlin.de/

Abstract. There is a growing movement around exploiting open data in in the web community. Applications publishing and managing public data of governments and administrations are becoming common, but use cases of open data in the business community are rare, and we have yet to demonstrate the benefit of open linked data in enterprise information management. In this paper, we present a business-related application that exploits open-data. "Active Hiring" is a search based application providing analytics on on-line job posts. This application uses services from the LOD cloud to disambiguate, geotag and interlink data entities acquired from on-line job boards web sites and provides a demonstration of the usefulness of linked open data in business setting.

Keywords: Search Based Application, Mashup, Geotagging, Analytics, Linked Data

1 Introduction

In recent months, a remarkable growth in the trend towards open data usage can be observed in the web community. Many applications have been developed for revealing and crossing open data concerning public data coming from government sources. The promise of open data has also lead commercial groups to consider how this linked open data paradigm can be used to redefine existing approached to business data integration. In this paper, we present "Active Hiring", a search based application providing analytics on on-line job posts. Active Hiring application uses services from the linked open data (LOD) cloud to disambiguate, geotag and link together key data entities extracted from crawled job boards web sites. Active Hiring is publicly accessible at the following location: http://activehiring.labs.exalead.com

2 Active Hiring Overview

Active Hiring is a search based application that provides comprehensive analytics on trends in on-line hiring market, as an input to business intelligence (BI) process for providing market insights and hiring intelligence. The application monitors several job board websites and outputs BI-like dashboards that may reveal job leads and support decision making. Active Hiring uses LOD datasets and services in data analysis to augment and consolidate information extracted from the websites. These open datasets allow our application to disambiguate key entities of Active Hiring by interlinking data and therefore grouping similar entities. The additional information from the linked open data cloud enriches the search and navigation experience provided by the Active Hiring application. Active Hiring has a data workflow architecture (see figure 1) that has been implemented using proprietary platform (EXALEAD CloudView^{TM})and open source semantic components from the LOD2 stack [1].Active Hiring's workflow consists of 4 steps: Data acquisition, Storage, Semantic Pipeline Processing and Data search.

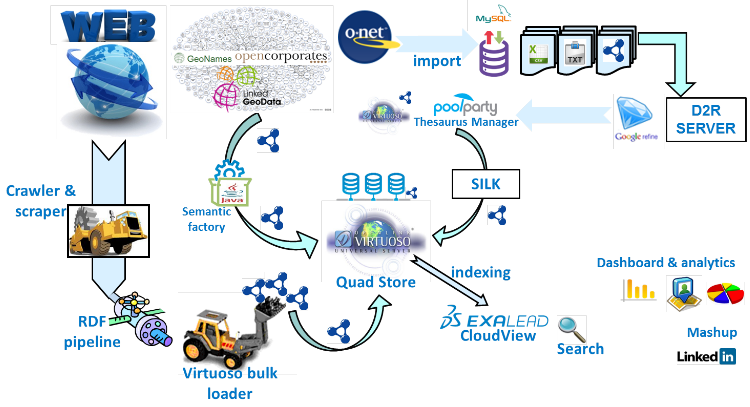


Fig. 1. Active Hiring architecture

2.1 Data Acquisition and Extraction

Once a list of job boards to monitor is selected, Active Hiring crawls the sites and extracts job posts. We used a semi-automatic scraping procedure for each monitored website to extract key properties of their job posts: job title, job location, hiring organization, job description, *etc.* This procedure is based on Scrapy [2], the open source scraping framework. The extracted properties are then streamed into a pipeline that transforms them into RDF triples conforming to the JobPosting class from schema.org [3].

2.2 Store

In a second step, Active Hiring uses the open source release of Virtuoso [4] quad store to cache the transformed crawled data as an RDF graph identified by an IRI[4]. With Virtuoso store, active hiring application provides a SPARQL endpoint that allows any user or remote service to access the crawled RDF triples via complex graph queries. It is accessible at the following location: `http://activehiring.labs.exalead.com/sparql`

2.3 Semantic Pipeline

Geotagging For each crawled job post, the semantic pipeline extracts the job location as a string value first. To find the precise geocoordinates of the detected named entity, a request is sent to LOD services, such as Geonames [5], for potential matches. The semantic pipeline uses then the disambiguation and interlinking procedure explained in the next section to select the most appropriate match. For each retained location match, the semantic pipeline fires a request to Geonames to get the geocoordinates of the considered location. These coordinates are appended to the considered job post by generating an RDF annotation, using the World Geodatic System (WGS84), which is inserted in the graph store. Below is an example of an annotation:

```
sparql
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX schema: <http://schema.org/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX geonames: <http://www.geonames.org/ontology#>
PREFIX wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#>
INSERT INTO <http://lod2.eu/wp8/active_hiring>
{
<LOC_URI> owl:sameAs <http://www.geonames.org/2968815>.
<http://www.geonames.org/2968815> wgs84:lat  "48.8534"^^xsd:string.
<http://www.geonames.org/2968815> wgs84:long  "2.3486"^^xsd:string.
<http://www.geonames.org/2968815> dc:name "Paris"^^xsd:string.
}
```

Disambiguation, reconciliation and interlinking This disambiguation module is one of the key features in the Active Hiring semantic pipeline. It attempts to remove potential ambiguity of company names and job locations references in job posts (*i.e.* values of "*hiringOrganization*" and "*jobLocation*" properties in "*JobPosting*" format). For example, the reference to "*Paris*" in a jobLocation property is ambiguous, i.e, it may correspond to (*Paris, Ile-De-France, France*) or to an American city, suc as (*Paris, Virginia, USA*). Similar ambiguities exist for company names. For example "*Dassault Systèmes*" may potentially match all the subsidiaries of the named company in the world.

We implemented a disambiguation/reconciliation procedure interlinking internal Active Hiring entities references with external entries from OpenCorporates [6] for company names and Geonames [5] entries for job locations, by using the country of a job location and the country of the hiring company to resolve ambiguity. The basic idea of our approach is that we use the country of candidate locations and the country of candidate companies to agree on the matching combination job location / hiring organization.

[4] Internationalized Resource Identifier

Vocabulary and taxonomy mapping As one might expect, job posts use a widely heterogeneous vocabulary to name the same job positions. For example, there are several variations when advertising for a "java developer" position: "java programer", "java engineer", *etc.* To link and cluster posts, we need to harmonize this vocabulary into a single business concept hierarchy.
In Active Hiring, we employ O*NET [8] and ESCO[5] taxonomies to harmonize job titles within the indexed data in our application. The O*NET taxonomy, released as a MySQL dump file, was transformed for Active Hiring into SKOS format using the D2R Server [9]. O*NET and ESCO generated SKOS files were then uploaded into PoolParty [10], a taxonomy management system and editor. To map job postings titles with our taxonomies' entries, we used the SILK [11] link discovery framework provided by the LOD2 Stack, applying the fuzzy matching method to extend the mapping to job titles that did not exactly match the taxonomies labels.

2.4 Search Application, Mashup and dashboardings

Active Hiring indexes the triples stored in the Quad store using a JDBC connection. Triples are mapped into a flat data model (key,values) stored in an inverted list index structure. Using this index structure, our application Active Hiring uses EXALEAD's CloudView platform [12] to provide search capabilities over the job vacancies triples content (i.e properties values). In addition to search, the Active Hiring homepage, see figure 2, provides an aggregation of dashboards that presents various business analytics computed on the indexed data.

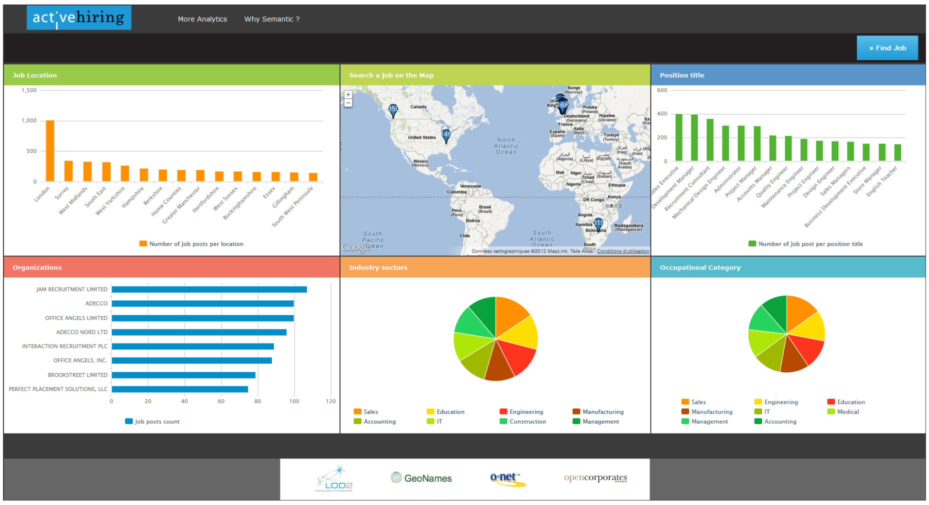


Fig. 2. Active Hiring Homepage Interface

[5] European Skills, Competencies and Occupations. ESCO is published in RDF SKOS

The main presented dashboards in Active Hiring interface are the following:

- Figure 3: a pie chart representing the distribution of job posts by industry sector. This widget provides an overview of the sectors that are actively recruiting.

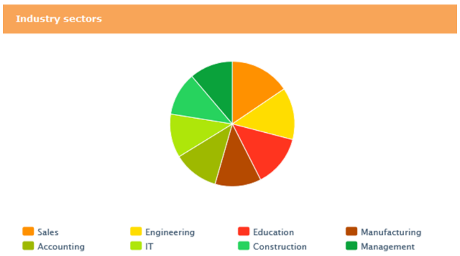


Fig. 3. Job posts distribution by industry sectors

- Figure 4: a world map presenting pinpoints of the indexed job posts by their location. The pinpoints corresponds to the geo coordinates that our semantic pipeline added to each job post. This widget visually highlights the geographical regions that have a dynamic recruitment market. We also provided a stacked columned chart view of the same data.

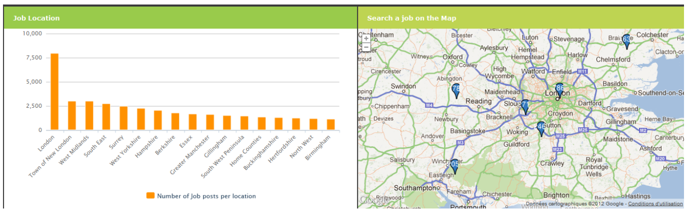


Fig. 4. Job posts distribution by locations

- Figure 5: Representation of the distribution of job posts by job titles. These widgets provide a view of the most demanded job positions.

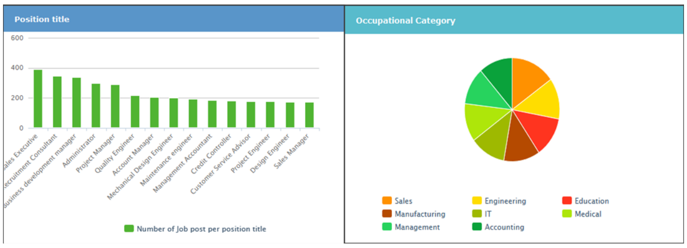


Fig. 5. Job posts distribution by position titles

- Figure 6: Representation of job posts distribution by hiring companies. This view highlights dynamic recruiting companies.

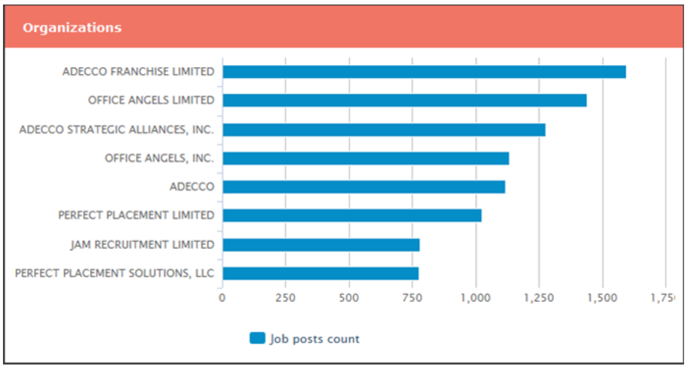
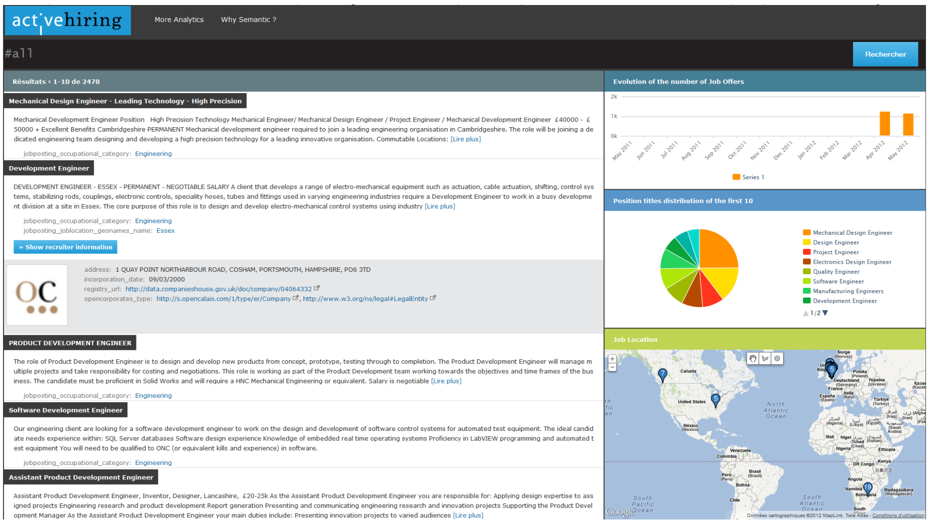


Fig. 6. Job posts distribution by hiring organization

the Active Hiring application also provides advanced faceting search to focus the query keywords on a specific field of a job posting. We defined 3 facets:

1. **job**: to focus the search on job position titles. Example: *"job:Sales Manager"*.
2. **loc**: to focus the search on job locations. Example: *"loc: london"*.
3. **industry**:to focus the search on industry domains. Example: *"industry:manufacturing"*

An auto-suggest facility specifc to each facet automatically displays indexed values to the user according to the first letters typed. Finally, Active Hiring



Mechanical Design Engineer - Leading Technology - High Precision

Mechanical Development Engineer Position High Precision Technology Mechanical Engineer/ Mechanical Design Engineer / Project Engineer / Mechanical Development Engineer £40000 - £ 50000 + Excellent Benefits Cambridgeshire PERMANENT Mechanical development engineer required to join a leading engineering organisation in Cambridgeshire. The role will be joining a dedicated engineering team designing and developing a high precision technology for a leading innovative organisation. Commutable Locations: [Lire plus]

jobposting_occupational_category: Engineering

Development Engineer

DEVELOPMENT ENGINEER - ESSEX - PERMANENT - NEGOTIABLE SALARY A client that develops a range of electro-mechanical equipment such as actuation, cable actuation, shifting, control systems, stabilizing rods, couplings, electronic controls, speciality hoses, tubes and fittings used in varying engineering industries require a Development Engineer to work in a busy development division at a site in Essex. The core purpose of this role is to design and develop electro-mechanical control systems using industry [Lire plus]

jobposting_occupational_category: Engineering
jobposting_joblocation_geonames_name: Essex

» Show recruiter information

address: 1 QUAY POINT NORTHARBOUR ROAD, COSHAM, PORTSMOUTH, HAMPSHIRE, PO6 3TD
incorporation_date: 09/03/2000
registry_url: http://data.companieshouse.gov.uk/doc/company/04064332
opencorporates_type: http://s.opencalais.com/1/type/er/Company, http://www.w3.org/ns/legal#LegalEntity

PRODUCT DEVELOPMENT ENGINEER

The role of Product Development Engineer is to design and develop new products from concept, prototype, testing through to completion. The Product Development Engineer will manage multiple projects and take responsibility for costing and negotiations. This role is working as part of the Product Development team working towards the objectives and time frames of the business. The candidate must be proficient in Solid Works and will require a HNC Mechanical Engineering or equivalent. Salary is negotiable [Lire plus]

jobposting_occupational_category: Engineering

Software Development Engineer

Our engineering client are looking for a software development engineer to work on the design and development of software control systems for automated test equipment. The ideal candidate needs experience within: SQL Server databases Software design experience Knowledge of embedded real time operating systems Proficiency in LabVIEW programming and automated test equipment You will need to be qualified to ONC (or equivalent kills and experience) in software.

jobposting_occupational_category: Engineering

Assistant Product Development Engineer

Assistant Product Development Engineer, Inventor, Designer, Lancashire, £20-25k As the Assistant Product Development Engineer you are responsible for: Applying design expertise to assigned projects Engineering research and product development Report generation Presenting and communicating engineering research and innovation projects Supporting the Product Development Manager As the Assistant Product Development Engineer your main duties include: Presenting innovation projects to varied audiences [Lire plus]

Fig. 7. Mashed content from OpenCorporates with query answers

3 Conclusion & Future Work

In this paper, we presented Active Hiring, a search based application that combines semantic technologies and services to produce Human Resources (HR) analytics and highlight major trends on on-line hiring market. This Active Hiring application is a demonstration of the benefit of combining open data sets and services with semantic tools as a support technology for increasing the accuracy of business applications.

Acknowledgments. The Active Hiring demonstrator has been developed within the activities of the European project LOD2 (see http://lod2.eu/).

References

1. LOD2: Lod2 stack. http://stack.lod2.eu/
2. Scrapy: An open source web scraping framework for python. http://scrapy.org/
3. Schema.org: http://schema.org/
4. Openlink_Software: Virtuoso universal server. http://virtuoso.openlinksw.com/
5. GeoNames: http://www.geonames.org/
6. OpenCorporates: http://opencorporates.com/
7. OpenCorporates_API: http://api.opencorporates.com/documentation/Home
8. O*NET: O*net online. http://www.onetonline.org/
9. Bizer, C.: D2r server. http://d2rq.org/d2r-server
10. PoolParty: Semantic information managment. http://poolparty.punkt.at/
11. Robert Isele, Anja Jentzsch, C.B.J.V.: Silk a link discovery framework for the web of data. http://www4.wiwiss.fu-berlin.de/bizer/silk/
12. EXALEAD: http://www.3ds.com/fr/products/exalead/

Appendix

3.1 Mandatory criteria

Targeted Audience Active Hiring targets Human Resources (HR) decision makers, providing dashboards, analytics and data views on on-line job market advertisements to support their daily tasks. End users simply looking for jobs can also use Active Hiring search interface.

Data sources The Active Hiring demonstrator uses several data sources: job posts from online job boards; data served by LOD services such as Geonames and OpenCorporates; HR taxonomies provided by O*NET and the European Commission. It involves a data ecosystem where several ownerships coexist: open data and owned data by job boards. This ecosystem aggregates data published in heterogeneous formats: MySQL dump for O*NET taxonomy, HTML for job posts, JSON/XML data for OpenCorporates and Geonames, etc. In this project, we processed and unified these formats into a standardized semantic RDF format.

Real world data Active Hiring acquired real world data from web sources (job boards, geographical datasets services, companies registries data) by crawling sources for one week, resulting in over 90,000 index job posts.

Meaning of data The Active Hiring semantic pipeline implements an interlinking approach based on disambiguating and clustering the processed data. For each processed job posts in Active Hiring workflow, our interlinking procedure disambiguates the combination ¡hiring company name,job location¿. We use several RDF/OWL formats to annotate the processed data and represent the interlinking and disambiguation output.

3.2 Desirable criteria

Web interface Active Hiring demonstrator provides a convivial web interface that uses asynchronous AJAX[6] calls to load content and update the rendering according to user interactions.

Scalability Active Hiring application uses EXALEAD CloudView™ to build its search based application and OpenLink Virtuoso server to manage RDF triples. EXALEAD CloudView™ is a scalable platform, for indexing and real time data search, that currently powers the general purpose EXALEAD[7] web search engine, indexing more than 16 billions of web pages. OpenLink Virtuoso Server is a scalable quad store that is used in production in many applications (DBpedia for example). In our demonstrator, we expose $90,000$ web job posts and more than 4 millions of triples. This amount can be extended without prejudice to the application or its response time.

Novelty Active Hiring is a innovative approach in using open data as a support technology for data integration and data analytics applications.

Functionality Active Hiring demonstrator is a search based application providing advanced analytics and data mashup features. 1

Commercial potential Active Hiring demonstrator is developed as a proof of concept. Its objective is to clearly target HR decision makers, job market surveyors and administrations since we believe in its commercial potential.

[6] Asynchronous JavaScript and XML

[7] `http://www.exalead.com/search`