

BPM in German Companies – Information Gathering

Felix Baumann and Dieter Roller

Institut für Rechnergestützte Ingenieursysteme, Universität Stuttgart

Abstract This position paper outlines a method devised to gather information on companies involved in BPM from publicly available data. The method is based on searching for job openings on job search engines which are related to BPM. The assumption is that companies searching for new employees with skills regarding BPM are either involved in BPM or are about to be involved in BPM.

Keywords: BPM, BPM in Germany, Crawler, Publicly Available Data, Web Scraping

1 Motivation

The statements made in the literature on the use of BPM in businesses are mainly based on statements which are provided by companies or their employees in the form of interviews and surveys [3] [1]. These surveys have partially the shortcoming that the company would like to represent itself positively or employees give answers in the surveys in accordance to the social desirability response set. Our idea is to identify companies involved in BPM by using their search for new employees on job sites. These sites are public as these companies need to address potential future employees. We can use this knowledge of the searching companies to make assertion on the prevalence of BPM within the German economy. For this purpose, publicly available sources, particularly job search engines are used to gather this information employin web scraping technology[5],[2]. False positive results can be excluded because companies will not search actively for new employees on topics/areas that are irrelevant for the organisation. There might be a bias in this case because we might miss some companies searching for employees when they don't use the correct keywords in their job descriptions when searching for the appropriate staff. After we have identified the companies/organisations that apply BPM or are about to apply BPM we can use the job advertisements itself in order to gain information on the necessary competencies/skills and related fields of work. Additionally information on the duration of the existence of each job advertisement can lead to information on the availability of BPM experts. In a further step, it is possible to correlate the firms found with publicly available information (self-representation on the Internet and corresponding sites like XING) and to draw conclusions about the geographical distribution of the companies, its business and number of employees. Thus it would be possible to support previous research results regarding the use of BPM, such as [7], [6], [4].

2 Technical Base

This work is based on the automated search and indexation of job advertisement with relation to BPM. As the sources of information the homepage of the Bundesagentur für Arbeit ¹ - with currently 829516 listed open positions (as of 2015-01-27) - and the German homepage of Stepstone ² - with currently 56398 listed open positions (as of 2015-01-27) - were chosen. The homepage of the Bundesagentur für Arbeit has been chosen because it is the governmental institution handling unemployed persons or persons currently searching for a job. Furthermore this site integrates other job search engines within its own search engine and makes those data sources available through a single interface. Stepstone has been chosen as a data source to counterbalance the governmental page of the JobBoerse as a privately owned entity. Stepstone is by its own accord Germanys job search engine number 1 (“Deutschlands Jobbörse Nr.1”). Other job search engines e.g. the German homepage of Monster³ have not been used because they either did give no account on the number of jobs they are listing or are available through the search engine of the JobBoerse - e.g. XING⁴ - or would only give replicates of jobs found on the two search engines used.

2.1 Research questions

We have identified the following set of questions that we want to answer.

1. Questions related to enterprises: (a) Is this method valid to acquire data on enterprises employing BPM? (b) Which enterprises do use BPM? (c) What kind of companies are these? (d) Where are they geographically located? (e) What is the size of these companies? (f) In which domain(s) are these companies located?
2. Questions related to the qualification of BPM practitioners: (a) What are qualities applicants are expected to have? (b) Is there an overlap between BPM and other fields of work like cloud computing? (c) What positions are BPM practitioners employed in? (d) What is the level of experience required? (e) Is the required skillset changing? (f) Do enterprises look for the same qualities as described in scientific literature[8]?
3. General questions: (a) What is the duration of each job advertisement? This could be indicative of abundance of of qualified employees? (b) Are the job openings recurrent (for the seemingly same position)? (c) Is there a general trend of enterprises employing BPM?

2.2 Implementation

Starting with the analysis of the datasources, we have devised a crawler able to search those datasources for a given keyword using the search functionality

¹ <https://jobboerse.arbeitsagentur.de>

² <http://www.stepstone.de>

³ <http://www.monster.de>

⁴ <http://www.xing.de>

provided by said sites. The data crawler consists of scripts which split the work into the following parts

1. Form the search query
2. Filter the results accordingly to the searched for information
3. Gather further information, e.g. URL of employer
4. Save the results in a persistent way

The crawler uses a variety of Linux software in order to fulfill the specific tasks (e.g. `wget`⁵ and `curl`⁶) and to filter the results by using streamline processing e.g. `awk`⁷ and `sed`⁸. The respective HTML pages are stored for documentation purposes twofold. They are stored as HTML files and audit-proof as PDF. The information extracted are stored in corresponding XML files and additionally in a SQLite⁹ database.

2.3 Data Acquisition

The actual keywords/phrases are being sent to the respective data sources/search engines daily. The crawler handles the communication with the search engines and stores the data persistently. Data collection has started 2014-09-18 and is ongoing. Currently there is 131 days worth of data available. There are 9 search queries with 9 keywords/phrases going to both search engines. The actual search phrases in use are: 1. “yawl” 2. “epk” 3. “epc” 4. “business process management” 5. “prozess modell” 6. “prozessmanager” 7. “sysml” 8. “bpmn” 9. “bpm”

2.4 Preliminary Data Analysis

The current data set consists of 767314 HTML pages out of which 104791 (13%) are originating from JobBoerse, 289065 (37%) from Stepstone and another 204737 (26%) pages from third parties accessed through the JobBoerse search engine. The remaining 24% of the pages is broken or otherwise impaired. These HTML pages have not been filtered for duplicates as of yet but there are certainly double entries that will be filtered out. Searching for unique job titles yields 7231 results. On average there seems to be 55 job openings per day spread over 2 search engines and 9 different keywords/phrases. Taking those different keywords and the different search engines into account it results in 3.06 job openings per day per keyword gathered.

2.5 Challenges

After the preliminary analysis the following challenges have arisen:

⁵ <https://www.gnu.org/s/wget>

⁶ <http://curl.haxx.se>

⁷ <http://www.gnu.org/s/gawk>

⁸ <https://www.gnu.org/software/sed>

⁹ <http://www.sqlite.org>

- Job openings by consultancy companies. Job advertisements by these companies can not necessarily be correlated to a specific enterprise which does the actual BPM work. Employees might be loaned to other companies or work as consultants for other companies. These job advertisements are identifiable by the company name and its classification as a consultancy company.
- Variations of the labour market. There might be other factors e.g. crises, wars or weather influencing the overall labour market. Taking other search phrases from other areas of information technology related areas into the data collection can reduce the effect of these variations as they can function as a baseline. Suggested terms include 1. cloud computing 2. data mining 3. java (programmer) 4. service engineer .

3 Conclusion and Outlook

The main contribution of our work will be an extensive set of data relating to companies involved in BPM, especially their geographic distribution, company size, domain and business. Furthermore the contribution will be a dataset of skills and related areas of business of people practising BPM in the real world. This data can be extend where necessary and used to support existing research in BPM using surveys and interviews based on participant selection. This is an ongoing experiment and we will be able to produce results without the need to halt data collection. A longer duration of data collection is necessary to track variations and trends in BPM and the associated job advertisements.

References

1. Cebeci, C., Kol, E.: Analysis for the Implementation of the Business Process Management in Selected Turkish Enterprises. *International Journal of Economics and Financial Issues* 3(2), 420–425 (2013)
2. Herrouz, A., Khentout, C., Djoudi, M.: Overview of Web Content Mining Tools. *The International Journal Of Engineering And Science (IJES)* 2, 106–110 (7 2013), <http://adsabs.harvard.edu/abs/2013arXiv1307.1024H>
3. Imanipour, N., Talebi, K., Rezazadeh, S.: Obstacles in Business process Management implementation and adoption in SMEs (1 2012), <http://ssrn.com/abstract=1990609>
4. Knuppertz, T., Schnägelberger, S., Clauberg, K.: Umfrage Status Quo Prozessmanagement 2010/2011. Tech. rep. (2011), http://www.bpmo.de/bpmo/export/sites/default/de/know_how/content/downloads/Status_Quo_Prozessmanagement_2011.pdf
5. Marres, N., Weltevrede, E.: Scraping the Social? Issues in live social research. *Journal of Cultural Economy* 6, 313–335 (4 2013), <http://dx.doi.org/10.1080/17530350.2013.772070>
6. Müller, T., Thome, R., Vogeler, K.: Zukunftsthema Geschäftsprozessmanagement. Tech. rep. (2011), <http://www.pwc.de/de/prozessoptimierung>
7. Wolf, C., Harmon, P.: The State of Business Process Management 2012. Tech. rep. (2012), <http://www.bptrends.com/bpt/wp-content/surveys/2012-BPT%20SURVEY-3-12-12-CW-PH.pdf>

8. Xu, L.D.: Enterprise Systems: State-of-the-Art and Future Trends. In: IEEE Transactions on Industrial Informatics. vol. 7, pp. 630–640. IEEE (09 2011), <http://dx.doi.org/10.1109/TII.2011.2167156>