Who-Does-What: A Knowledge Base of People's Occupations and Job Activities

Jonas Bulegon Gassen[1], Stefano Faralli[2], Simone P. Ponzetto[2], and Jan Mendling[1]

[1] Vienna University of Economics and Business
Augasse 2-6, A-1090 Vienna, Austria
{jonas.gassen,jan.mendling}@wu.ac.at,
[2] Data and Web Science Group, University of Mannheim, Germany
{stefano,simone}@informatik.uni-mannheim.de

Abstract. We present a novel resource called "Who-Does-What" (WDW), which provides a knowledge base of activities for classes of people engaged in a wide range of different occupations. WDW is semi-automatically created by automatically extracting structured job activity descriptions from the Web (we use here the O*Net website). These descriptions are used to populate the taxonomic backbone provided by the manually-created Standard Occupational Classification (SOC) of the US Department of Labor.

1 Introduction

System analysis and design is concerned with the creation of conceptual models of various aspects of a discourse domain. One of its key challenges is quality assurance, in particular regarding elements' labels [2, 8]. This could be addressed, for instance, by leveraging techniques for automatic recommendation of (e.g., activity) labels during modeling [7]: however, to date, there exists no specific knowledge resource that could potentially enable knowledge-rich and domain-specific recommendation techniques for conceptual modeling, e.g., by providing wide-coverage structured knowledge about subjects (i.e., actors), typical verbs (i.e., actions) and corresponding objects.

In this paper, we set to fill this gap and describe a novel resource called "Who-Does-What" (WDW) that organizes knowledge on activities and the classes of people that typically perform them. We connect classes of people to a wide range of different occupations, like computer programmers or bakers. WDW is semi-automatically created by populating the manually-created taxonomy from the Standard Occupational Classification (SOC) of the US Department of Labor with activities found in the web. We extract activities (i.e., predicates and their arguments) and automatically acquire the job duties related to each occupation. These structured representations of activities are linked to the backbone taxonomy (SOC).

Our resource is meant as a first step towards the more general goal of ontology-rich semantic modeling: here, we focus on the important task of extracting occupation-related activities from text, linking them to a taxonomy and representing such knowledge explicitly in a clean semantic form. Previous work covered related tasks such as automatically extracting occupation-related concepts from text for the task of mining

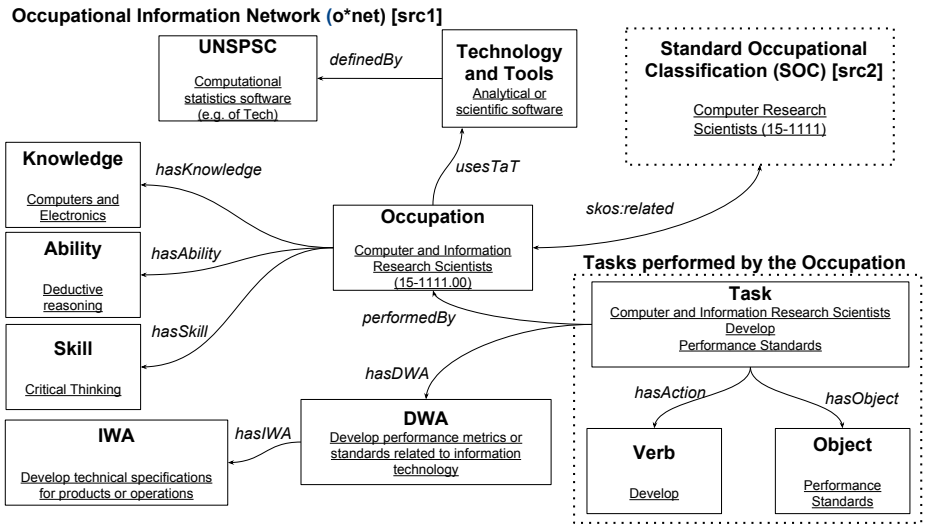

Fig. 1. Schema of WDW (version 1.0)

biographic information [6] and developing ontologies to support experts finding systems [1]. Our approach can be seen as a specific case of the more general task of ontology population from text [11]: it leverages techniques previously developed in the context of open information extraction systems like NELL [3] or ClausIE [4], which have been shown to be capable of acquiring large amounts of machine-readable knowledge from text, which can later be linked to wide-coverage ontologies [5]. Our long-term vision is to effectively support, among other tasks, label completion of process models based on such resources. This is listed as number 5 among the 25 challenges of semantic process modeling [9]. In this way, it can help improving label quality of process models [8], class diagrams [2] and other types of conceptual models [10].

2 A knowledge base of people's occupations and job activities

We present the schema of WDW in Figure 1 and describe the approach used to build our resource. We use two sources: the Standard Occupational Classification (SOC) and O*Net[3]. The schema labels are bold text and one instance example is shown with underlined text. We connect SOC to their related occupations on O*Net based on additional information about occupations such as skills or knowledge. Additionally, we generate tasks in a triple format, connecting them to occupations and to information from O*Net, namely Detailed Work Activity (DWA) and Intermediate Work Activity (IWA):

1. **Taxonomy harvesting**. We make use of the manually-created taxonomy from the Standard Occupational Classification (SOC) of the US Department of Labor as a taxonomic backbone for WDW.

[3] `http://www.bls.gov/soc/` and `http://www.onetcenter.org/`

2. **Additional information about occupations**. We collect additional information related to occupations of SOC: technology and tools that are used by the occupations and knowledge, abilities and skills that might be required in such occupations. The source for this information is O*Net.
3. **Extracting propositions of occupation-related activities**. We assemble textually represented tasks from O*Net. Given the relevant text fragments, we apply a state-of-the-art Open Information Extraction system to turn the semi-structured activity description into structured representations.

Resource deployment. To share an RDF/OWL version of our resource we create an OWL representation of the SOC ontology, extend the O*Net schema with 3 new tables for "task", "verb" and "object", generate an RDF file with D2RQ and map all occupations from O*Net to the SOC ontology. All data are freely available under a CC BY-NC-SA 3.0 license at `https://madata.bib.uni-mannheim.de/179/`.

Using SOC as taxonomic backbone. We use the manually-built Standard Occupational Classification (SOC) of the US Department of Labor (current version from 2010) as backbone taxonomy for our resource. As unique identifiers, we use the label of the occupation descriptions concatenated with the SOC code.

Codes in the SOC hierarchy are made up of six digits divided by a hyphen, e.g. 51-3011 refers to the class BAKERS.The first two digits represent the top-level class (51-0000: Production Occupations), whereas the third digit represents the mid-level class (51-3000: Food Processing Workers). The fourth and fifth digits represent the broad occupation (51-3010: Bakers) and the sixth digit represents the detailed occupation. Each occupation has a description and specific examples, like, Bread Baker or Bagel Maker. The full SOC hierarchy tree contains 1,421 occupation classes. In the OWL file, all occupations contain an rdfs:comment "SOCID". The leaves of the hierarchy contain examples of job titles and a textual description, both as rdfs:comment. There are 23 top-level classes, all branches have the maximum depth of 4, as for SOC code.

Collecting sentences describing job activities from the SOC hierarchy. We harvest information from O*NET OnLine website[4], which provides us with semi-structured descriptions of the SOC's concepts. We used the table "Tasks to DWAs" from O*Net database. As for now, we used only the text from Tasks because they appear to be more specific, e.g., for "bakers":

1. **Task:** Check products for quality and identify damaged or expired goods;
   - **DWA:** Evaluate quality of food ingredients or prepared foods;
     - **IWA:** Evaluate production inputs or outputs.
2. **Task:** Set oven temperatures and place items into hot ovens for baking;
   - **DWA:** Adjust temperature controls of ovens or other heating equipment;
     - **IWA:** Adjust equipment to ensure adequate performance.

[4] `http://www.onetonline.org`

Extracting structured job activity descriptions. We use the state-of-the-art Open Information Extraction system ClausIE [4] to process our sentences and extract structured triples from them. To maintain high precision across the output extractions we make use of simple heuristics for filtering: i) we keep only triples whose objects contain one or two words, each of at least three characters; ii) we remove triples where the predicate is a verb that is either auxiliary, modal or intransitive that cannot be used transitively (such verbs are detected based on blacklists created using Wiktionary). These verbs are removed because they are unlikely to be used in standard conceptual models, e.g., capturing business processes. ClausIE may retrieve triples with modal verb such as *has* or *might*, which do not denote an action as expected in process models. As a result, we obtain 5,548 triples from the O*NET corpus.

3 Conclusions

We presented Who-Does-What (WDW), a knowledge base of people's occupations and job activities. Our resource is a first step towards the more general goal of increasing the quality of conceptual models' labels, e.g. by enabling knowledge-rich automatic completion and recommendation techniques for semantic process modeling. WDW is freely available under a CC BY-NC-SA 3.0 license at `https://madata.bib.uni-mannheim.de/179/`.

References

1. W. Abramowicz, E. Bukowska, J. Dzikowski, A. Filipowska, and M. Kaczmarek. Semantically enabled experts finding system – ontologies, reasoning approach and web interface design. In *ADBIS*, pages 157–166, 2011.
2. D. Aguilera, C. Gómez, and A. Olivé. A complete set of guidelines for naming UML conceptual schema elements. *Data & Knowledge Engineering*, 88:60–74, 2013.
3. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, pages 1306–1313, 2010.
4. L. D. Corro and R. Gemulla. ClausIE: clause-based open information extraction. In *WWW*, pages 355–366, 2013.
5. A. Dutta, C. Meilicke, and S. P. Ponzetto. A probabilistic approach for integrating heterogeneous knowledge sources. In *ESWC*, pages 286–301, 2014.
6. E. Filatova and J. M. Prager. Occupation inference through detection and classification of biographical activities. *Data & Knowledge Engineering*, 76:39–57, 2012.
7. A. Koschmider, T. Hornung, and A. Oberweis. Recommendation-based editor for business process modeling. *Data & Knowledge Engineering*, 70(6):483–503, 2011.
8. H. Leopold, J. Mendling, and O. Gunther. What we can learn from quality issues of BPMN Models from industry. *IEEE Software*, 2016.
9. J. Mendling, H. Leopold, and F. Pittke. 25 challenges of semantic process modeling. *IJISEBC*, 1(1):78–94, 2014.
10. F. Pittke, B. Nagel, G. Engels, and J. Mendling. Linguistic consistency of goal models. In *BPMDS*, pages 393–407, 2014.
11. W. Wong, W. Liu, and M. Bennamoun. Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 44(4):20:1–20:36, 2012.