A Linked Data Approach to Know-How

Paolo Pareti[1,2], Benoit Testu[1], Ryutaro Ichise[1],
Ewan Klein[2], and Adam Barker[3]

[1] National Institute of Informatics, Tokyo, Japan,
p.pareti@sms.ed.ac.uk, benoit.testu@u-psud.fr, ichise@nii.ac.jp
[2] University of Edinburgh, Edinburgh, United Kingdom
ewan@inf.ed.ac.uk
[3] University of St Andrews, St Andrews, United Kingdom
adam.barker@st-andrews.ac.uk

Abstract. The Web is one of the major repositories of human generated know-how, such as step-by-step videos and instructions. This knowledge can be potentially reused in a wide variety of applications, but it currently suffers from a lack of structure and isolation from related knowledge. To overcome these challenges we have developed a Linked Data framework which can automate the extraction of know-how from existing Web resources and generate links to related knowledge on the Linked Data Cloud. We have implemented our framework and used it to extract a Linked Data representation of two of the largest know-how repositories on the Web. We demonstrate two possible uses of the resulting dataset of real-world know-how. Firstly, we use this dataset within a Web application to offer an integrated visualization of distributed know-how resources. Lastly, we show the potential of this dataset for inferring common sense knowledge about tasks.

1 Introduction

Cooking recipes, software tutorials and standard operating procedures are some examples of the procedural knowledge currently available on the Web. We call this particular type of knowledge *human know-how*. Unlike other well-known types of procedural knowledge, such as program code, human know-how describes procedures which are primarily meant to be executed by humans.

Human know-how on the Web could be used in a wide variety of applications, such as for Information Retrieval and Service Recommendation [2]. Intelligent systems accessing this knowledge can also directly benefit Web users by discovering relevant procedures and resources. These potential applications, however, are currently hindered by the limitations of the available know-how. The most severe limitations are the lack of structure and the isolation from other knowledge sources. To overcome these limitations, we have developed a Linked Data framework which can represent human know-how. This framework can automatically acquire the Linked Data representation of procedures from existing Web articles and then integrate it with related resources.

Existing methods for representing procedural knowledge are usually concerned with a moderate number of well defined procedures. Human know-how on the Web, on the contrary, is inherently noisy, in constant evolution, large in size, distributed across multiple repositories and covers many different domains. The proposed Linked Data representation is robust to these challenges and, where needed, can act like a bridge to more sophisticated representations.

Our knowledge extraction and integration experiments resulted in the creation of a large amount of Linked Data representing real-world human know-how. We present two possible uses of this dataset. Firstly, an integrated visualization of distributed resources through a Web application. Lastly, examples of semantically rich queries that can extract common sense knowledge from this dataset.

1.1 Methodology

Our framework for the representation of human know-how is divided into two components. The first component automates the extraction of know-how from existing Web articles. The extracted knowledge is represented using a simple and light-weight Linked Data vocabulary [3]. Similarly to the DBpedia project [1], we focus our extraction on semi-structured resources. Semi-structured know-how resources, such as the articles available on WikiHow,[4] can be analyzed reliably, as they are explicitly divided into a number of steps, methods and requirements.

The second component of our framework automates the integration of the extracted know-how with other related knowledge. One possible integration involves the inputs and the outputs of a procedure, which can be linked to the related DBpedia types. For example, the ingredient of a procedure labelled "30 grams of sugar" can be linked to the DBpedia entity representing the concept of sugar. Links can also directly connect different procedures. For example, the step "prepare a short resume" could be linked to the relevant procedure "how to prepare a resume". The generation of those links is based on a Machine Learning classifier which utilizes, among others, textual features extracted using Natural Language Processing.

1.2 Experimental Results

We have implemented our framework and we have validated it in a real world setting. Our experiment resulted in the extraction of over 200,000 procedures from two different know-how websites, namely WikiHow and Snapguide.[5] These procedures were represented as Linked Data and integrated with (1) other know-how resources and (2) other existing Linked Data, such as DBpedia. We have evaluated the quality of our automatic integration against existing integration efforts and showed its superiority across all the dimensions considered, such as the quantity and the precision of the links. For a more detailed account of this framework see the forthcoming paper by P. Pareti et al. [4].

[4] `http://www.wikihow.com/`
[5] `http://snapguide.com/`

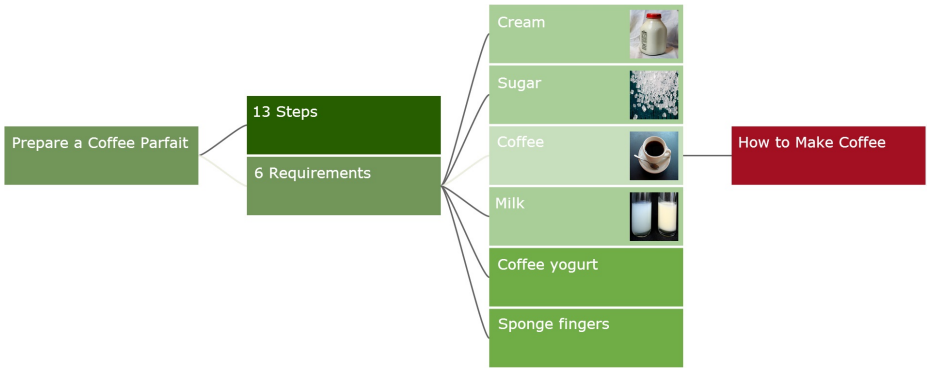

Fig. 1. Screenshot of the Web application

2 Case Studies

To better understand the challenges and the opportunities of a Linked Data representation of human know-how we will present two different uses of a real-world dataset. Firstly, we will describe a Web application which offers an integrated visualization of distributed know-how resources. Secondly, we will show the potential of this dataset for inferring common sense knowledge. The dataset used in these case studies is the one extracted by our system from the WikiHow and Snapguide websites. This dataset includes the links generated by our automatic integration system, such as the links between related WikiHow and Snapguide instructions.

2.1 Web Application

The results of our knowledge extraction and integration experiment are available online through a Web application.[6] This application allows users to visualize procedures in a tree-like structure. After finding a procedure using keyword search, the title of this procedure is displayed as the root node of the tree. By clicking on the children of this node, the user can expand the representation and obtain more information about the steps, the requirements and the outputs of this procedure. The links generated by our integration experiment allow the original tree to be expanded with additional information on the related procedures and DBpedia types. For example, Figure 1 shows a screenshot of this application visualizing the requirement "Coffee" of the procedure "Prepare a Coffee Parfait". The description of this requirement originally consisted only of a short textual label. Thanks to the links generated by our integration system, this description is now expanded with a picture of its type retrieved from DBpedia, and the link to a procedure that can create such requirement. This application demonstrates how

[6] `http://w3id.org/prohow/main/`

Linked Data can be used to have an integrated visualization of both procedural and declarative knowledge retrieved from different sources.

2.2 Semantic Queries

Human know-how can be analysed to infer common sense knowledge. This can be demonstrated by running semantically rich queries on the dataset generated by our experiment. One example of such queries can estimate the generality of a task. This can be used to distinguish very specific tasks, such as "how to install Android OS 4.3 on Windows 8", from basic and highly reusable tasks, such as instructions on "how to preheat an oven". Basic tasks can help identifying useful basic skills, and are ideal candidates for optimization and automation. Another example of a semantically rich query can evaluate the correlation between entities. For example, it is possible to discover the common sense fact that the entity "pen" and the entity "paper" are typically used together.

3 Conclusion

In this document we have introduced our Linked Data framework to represent human know-how on the Web. This framework has been tested in a knowledge extraction and integration experiment which generated a large dataset of real-world know-how. We have described two possible applications of this dataset. The first is a Web application for the visualization of distributed know-how resources. The second is the direct acquisition of common sense knowledge from the dataset using semantically rich queries.

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer Berlin Heidelberg, 2007.
2. S.-H. Myaeng, Y. Jeong, and Y. Jung. Experiential Knowledge Mining. *Foundations and Trends in Web Science*, 4(1):71–82, 2013.
3. P. Pareti, E. Klein, and A. Barker. A Semantic Web of Know-how: Linked Data for Community-centric Tasks. In *Proceedings of the 23rd International Conference on World Wide Web Companion*, pages 1011–1016, 2014.
4. P. Pareti, B. Testu, R. Ichise, E. Klein, and A. Barker. Integrating Know-How into the Linked Data Cloud. In *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management*, 2014. (forthcoming).