Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
An Adjunct to the Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing

# Crowdsourcing the Extraction of Data Practices from Privacy Policies

**Florian Schaub, Travis D. Breaux, Norman Sadeh**

School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
fschaub@cs.cmu.edu, breaux@cs.cmu.edu, sadeh@cs.cmu.edu

**Abstract**

Website and mobile application privacy policies are intended to describe the system's data practices. However, they are often written in non-standard formats and contain ambiguities that make it difficult for users to read and comprehend these documents. We propose a crowdsourcing approach to extract data practices from privacy policies to provide more concise and useable privacy notices to users and support the analysis of stated data practices. To that end, we designed a hierarchical task workflow for crowdsourcing the extraction of data practices from privacy policies. We discuss our workflow design and report preliminary results.

## Motivation

Privacy policies are the de facto standard in the U.S. and Europe for describing data practices of websites and mobile applications. However, privacy policies are complex text documents, because they use different formats, contain references to other policies and contain intended and unintended ambiguity. Most users do not read privacy policies (Cranor 2012), yet they implicitly consent to a service's privacy policy by using the service. We propose to use crowdsourcing to extract the data practices stated in privacy policies in order to make this information more accessible. We aim to improve accessibility by displaying concise privacy notices to users that highlight unexpected data practices (Sadeh et al. 2013), such as sharing contact or payment data with third parties.

Crowdsourcing the extraction of data practices faces several challenges: (1) a whole policy may contain over 100 sentences, which can require 15-30 minutes to read; and (2) it is difficult to ensure that workers read the whole policy to inform their responses. Because of these two challenges, a task that requires workers to read a whole policy places a large payment at risk of low quality work. By comparison, a task decomposition approach would distribute the risk across smaller subtasks that depend on answers from multiple workers to obtain sufficient annotation accuracy. We first review related work before presenting our results to create a scalable annotation workflow.



## Related Work

Privacy policy complexity is the focus of standardizing machine-readable alternatives, such as P3P, Do Not Track and improved notice formats (Cranor 2012). However, few have investigated how data practices can be extracted from privacy policy documents. Breaux et al. (2014) propose a full frame extraction approach for trained analysts. *Terms of Service; Didn't Read* (http://tosdr.org) is a community-based crowdsourcing approach, which relies on community members to contribute to company data practice profiles. Scalability of these efforts has not been the focus of their evaluation, however. Others investigate natural language processing (NLP) techniques to extract information from privacy policies (Ammar et al. 2012, Zimmeck and Bellovin 2014, Ramanath et al. 2014). Yet, these approaches overlook nuanced descriptions of data practices due to ambiguity or annotator disagreement. Based on our prior work (Breaux and Schaub 2014, Ammar et al. 2012) and inspired by work on crowdsourcing category extraction from complex data (André et al. 2014), we propose a hierarchical task workflow that enables crowdsourcing extraction of detailed data practices from privacy policy documents.

## Extracting Overview

Extracting data practices from policies entails four steps:

*1. Policy acquisition.* First, privacy policies are obtained from websites. This step can be crowdsourced or semi-automated using patterns for privacy policy URLs. Policies need to be re-acquired whenever the original policy is being updated. TOSback is an effort to monitor changes in privacy policies (https://tosback.org/).

*2. Policy segmentation.* The privacy policy is divided into small segments of about 120 words to reduce worker reading times. Segments must retain semantic coherence of contained statements (e.g., anaphora should match previously stated nouns and verbs in the same segment).

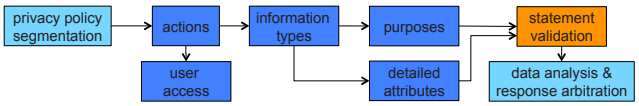


*Figure 1: Hierarchical task workflow with subtasks.*

*3. Data practice extraction.* Privacy policies typically contain various data practices (e.g. collection, third party sharing), which can be extracted from policy segments by formulating respective crowdsourcing tasks. Our task workflow design is presented in detail in the next section.

*4. Data analysis.* Data practices extracted from multiple policy segments need to be evaluated to determine a service's overall data practice posture. This involves reconciling contradictory data practices.

## Hierarchical Task Workflow

In step 3, the task workflow is based on a manual full frame extraction method (Breaux et al. 2014), which we adapted for crowdsourcing (Breaux and Schaub 2014). We present our current efforts on designing a hierarchical task workflow for such extraction tasks.

Figure 1 presents the decomposition: first we present a policy segment to workers and ask the worker to highlight action words (verbs) by pressing a keyboard shortcut that links the word to an action category (see Figure 2). The categories are *collection, use, transfer/sharing, retention,* and *consent.* Based on early pilot studies, our results show that 5–15 workers are required to stabilize agreement and that crowdworkers identified more action verbs in comparison to manual expert annotations (Breaux and Schaub, 2014). This task can be complemented with NLP to identify modal qualifiers (e.g., may or may not collect).

In the next task, we present segments with extracted action words highlighted and ask workers to highlight the relevant information types on which the action is performed. Related information types can be later grouped to achieve more concise data practice statements (e.g., e-mail address and phone number are contact information).

In the third step, we ask participants to identify purposes, e.g., the reasons for collecting contact information. Depending on the type of identified action, additional subtasks may be needed to extract the source of information collection, targets of information sharing, the user consent model, retention periods, and if users may modify data.

After collecting data via subtasks, we aim to recompose answers using templates to obtain data practices, e.g., the following template describes a collection practice where terms in brackets represent answers from subtasks:

> <*modality*> collection of <*information type*> from <*source*> for <*purpose*> with <*consent model*> stored for <*retention period*>.

We intend to use these instantiated templates in subsequent validation tasks shown to crowdworkers alongside the

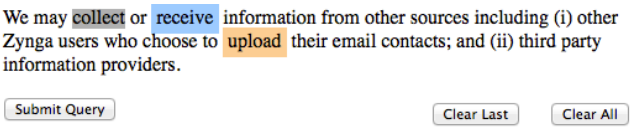


*Figure 2: Task UI for identification and categorization of action verbs in privacy policies.*

original policy segment. If workers indicate that the statement is incorrect, they can be asked to correct the statement, including the action or information type.

Further, we propose to use Description Logic to reason about these templates with the goal of drawing inferences from ambiguous and divergent policy interpretations across different policy segments (Breaux et al. 2014).

## Outlook

Herein, we described a comprehensive task workflow for crowdsourcing the extraction of data practices from privacy policies. We are presently evaluating and refining these subtasks. Initial results indicate a 1.5:1 cost reduction compared to extraction by two trained analysts (Breaux and Schaub, 2014), with the added benefit that the worker pool exceeds the number of available trained analysts. We intend to scale up this extraction process by incorporating NLP-based methods and refining our task decomposition.

This work is supported by NSF grant CNS 13-30596.

## References

Ammar, W., Wilson, S., Sadeh, N., Smith, N.A. 2012. *Automatic Categorization of Privacy Policies: A Pilot Study*. Tech report CMU-ISR-12-114, Carnegie Mellon University.

André, P., Kittur, A., Dow, S. P. 2014. *Crowd Synthesis: Extracting Categories and Clusters from Complex Data*. Proc. CSCW '14, ACM.

Breaux, T.D., Hibshi, H., Rao A. 2014. *Eddy, A Formal Language for Specifying and Analyzing Data Flow Specifications for Conflicting Privacy Requirements. Req'ts Engr. J.*

Breaux, T. D., and Schaub, F. 2014. *Scaling Requirements Extraction to the Crowd: Experiments with Privacy Policies.* Proc. Intl. Req'ts Engr. Conf. (RE '14). IEEE.

Cranor, L. F. 2012. *Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice*. J on Telecomm & High Tech. L., vol 10.

Ramanath, R., Liu, F., Sadeh, N., Smith, N. A. 2014. *Unsupervised alignment of privacy policies using hidden Markov models.* Proc. Association of Computational Linguistics (ACL '14).

Sadeh, N., Acquisti, A., Breaux, T. D., Cranor, L. F., McDonald, A. M., Reidenberg, J. R., Smith, N. A., Liu, F., Russell, N. C., Schaub, F., Wilson, S. 2013. *The Usable Privacy Policy Project.* Tech. report CMU-ISR-13-119, Carnegie Mellon University.

Zimmeck, S. and Bellovin, S. M. 2014. *Privee: An Architecture for Automatically Analyzing Web Privacy Policies.* Proc. USENIX Security Symposium. USENIX Association.