

PaperStream: Encoding Paper Surveys and Questionnaires Automatically

Julio Vega, Markel Vigo, Caroline Jay, Simon Harper

School of Computer Science, University of Manchester

julio.vega@manchester.ac.uk

Researchers from different disciplines use paper diaries and surveys to collect self-reported data about participants' experiences and opinions during a research study. Even though paper is a cheap, robust, and accessible method to collect data, encoding (transcribing) all that information into a digital file is time-consuming and error-prone.

PaperStream [1] is a web application to create paper surveys or questionnaires that can be printed and answered by participants using a regular pen. This software can then read and store their answers automatically into a digital file ready to be processed in Excel, R or SPSS. Researchers can forget about the arduous process of data collection and spend more time analysing it.

PaperStream has two key features. Creation of booklets, where it takes a one-page survey/questionnaire and generates a new file containing that template on each page labelled with the current date, participant ID and page number. This booklet is ready to print as a double-sided A4 or A5 document which can then be given to participants to collect data. Although PaperStream was created with self-reporting diaries in mind, researchers can use it with any multiple-answer questionnaire or survey. Secondly, this software can encode the marks on each page of the printed booklet into a CSV file. To do this, the researcher needs to scan a blank page of the deployed survey/questionnaire and through a web interface, they can mark all its relevant areas and their corresponding values which PaperStream will use as an encoding rubric (for example, identifying seven different blobs on the template as part of a pain scale). After the rubric is ready, the researcher just needs to scan the answered booklets used by their participants into TIFF or PNG images which will be encoded accordingly. Researchers can use the automatic feeder of a copying machine to scan all the required pages and do not need to worry about aligning them perfectly, PaperStream will identify the answer area automatically. In the resulting CSV, there will be one column per question and one row per page of the survey/questionnaire. Additionally, it will flag any missing or repeated data in case participants forgot to respond to a question or made any pen annotations.

PaperStream is open source [2] and written in Python. It uses image recognition algorithms to encode participants' answers and is packaged as a desktop app that can be used in Windows, Mac OS or Linux. It was deployed for the first time to collect self-reported symptom fluctuations of people with Parkinson's Disease and the booklets created with it implemented the five design recommendations for Parkinson's monitoring artefacts to be published in the ACM CHI Conference 2018 [3].

References:

1. Vega, J. PaperStream. <https://paperstream.netlify.com/>. Accessed February 1 2018.
2. Vega, J. JulioV/paperstream:PaperStream. Zenodo. November 13 2017. <http://doi.org/10.5281/zenodo.1048283>
3. Vega J., Couth S., Poliakoff E., Kotz S., Sullivan M., Jay C., Vigo M, and Harper S. Back to Analogue: Self-Reporting for Parkinson's Disease. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). <https://doi.org/10.1145/3173574.3173648>