Measuring and Improving Data Quality of Media Collections for Professional Tasks

Myriam C. Traub[*]
Centrum Wiskunde & Informatica
Science Park 123
Amsterdam, The Netherlands
firstname.lastname@cwi.nl

ABSTRACT

Carrying out research tasks on data collections is hampered, or even made impossible, by data quality issues of different types, such as incompleteness or inconsistency, and severity. We identify research tasks carried out by professional users of data collections that are hampered by inherent quality issues. We investigate what types of issues exist and how they influence these research tasks. To measure the quality perceived by professional users, we develop a quality metric. This allows us to measure the suitability of the data quality for a chosen user task. For a chosen task, we study how the data quality can be improved using crowdsourcing. We validate our quality metric by investigating whether professionals perform better on the chosen research task.

1. MOTIVATION

Digitization initiatives in numerous libraries and archives and (linked) open data projects lead to a growing amount of digital information that can be used for research. While some disciplines within the humanities, such as literary studies [3], have already adopted research questions and practices that make use of digital data, other disciplines are still at an earlier stage of this process. This paradigm shift caused researchers to reflect on the changes that are required in their approaches [1] and how the new practices can extend the current research landscape.

The data custodians, on the other side, put effort in making more content available in a way that users can easily access and navigate through it. The evaluation of digital archives and libraries needs to deal with a variety of aspects: data quality in respect to completeness, accuracy and consistency [6], usability of the interfaces and biases caused by selective digitization and collection policies [10]. On top of this, specific requirements of research tasks towards data enrichment and presentation have to be taken into account as different tasks may e.g. weigh precision and recall differently [11]. For some tasks, objectively measurable aspects are crucial, while for other tasks the subjective perspective of users is more important [7].

To our knowledge, no research has so far evaluated how well the data of digital archives supports *specific* research tasks of humanities researchers. Our research will therefore focus on the evaluation of data fitness for specific research tasks and how it can be improved.

To make sure that the data in libraries and archives meets the requirements of researchers, improvements would ideally be made by experts, such as archivists and librarians. Their expertise, however, is costly and the size of the data sets makes this approach unaffordable for the institutions. Training automatic methods would lead to high output, however, it will be difficult to obtain sufficient precision due to the heterogeneity of the data and the lack of expert judgments as training data. Crowdsourcing can be an efficient and effective way to get simple tasks performed by a large amount of people. For tasks with higher complexity or required expert knowledge, however, users must be trained in order to fully understand the task and provide high quality contributions. Our initial study [8] suggests that by combining these three components and creating feedback loops between them, we can create a system that successively leads to substantial improvements in data sets.

In order to measure the improvement of the data quality, we need a suitable quality metric. Using the data custodians' judgements as a quality measure for data may not always reflect the usefulness of the data for professional users. Therefore, we aim at developing a quality metric that allows us to measure how well the data suits the users' needs when performing a certain task. In this way, we can measure whether or not the crowdsourced contributions are actually perceived as an improvement and therefore considered useful.

2. RESEARCH QUESTIONS

Research shows that crowds are able to perform simple tasks (e.g. estimating the weight of an ox) with a precision that is close or even better than judgements given by experts of the field [5]. A more difficult task (judging biopsy images according to visual clues) has been crowdsourced by [4]. They showed that experts using the crowdsourced data can improve the precision of their diagnoses. To improve the quality of large data collections, we look into crowdsourcing tasks of a higher level of complexity.

We investigate how we can enable crowd workers to make

[*] Third year PhD student at CWI, supervised by Jacco van Ossenbruggen and Lynda Hardman.

IIiX '14 Aug 26-29 2014, Regensburg, Germany
Copyright is held by the owner/author(s).
ACM 978-1-4503-2976-7/14/08.
http://dx.doi.org/10.1145/2637002.2637056 .

contributions to a professional data set that are perceived as an improvement on quality by the professional users.

RQ1: Can crowd workers contribute data that is in line with expert contributions?

- **a.)** How do crowd workers performing a simplified expert classification task compare to experts?
- **b.)** Do crowd workers become better at performing the task and, if so, is that only on repeated items or also on new items?
- **c.)** How does the partial absence of the correct answer affect the performance of the crowd workers?

Social science researchers use digital data collections mainly in the explorative phase of their research. The approaches suggested in [2] therefore focus only on this research phase. We investigate what the requirements of this group of professional users towards the quality of the data would be for it to be useful in later phases of their research. With the help of humanities researchers we spot quality issues in large data collections and develop a quality metric that allows us to measure potential improvements.

RQ2: How do professional users perceive the effect of data quality on task execution?

- **a.)** Which tasks are affected by quality issues in data?
- **b.)** How do quality issues in data sources impact tasks performed by professional users of digital archives?
- **c.)** Which of these tasks are considered most important by professional users?
- **d.)** What is a suitable quality metric to measure the effect of data quality issues on tasks carried out by professional users?

Once we know what the main quality issues are and how they affect the work of professional users, we investigate how to crowdsource the improvements. We measure the usefulness of the contributions using a user-based quality metric.

RQ3: How can we apply crowdsourcing to improve the data quality as measured by our metric?

- **a.)** What is a suitable crowdsourcing task to improve the targeted quality issue?
- **b.)** How large is the gain in quality according to our previously defined user-based quality metric?

We validate the user-based quality metric by comparing results from a user study with the quality measured.

RQ4: Is the data gained through crowdsourcing useful for a professional user carrying out the chosen research task?

- **a.)** Can we validate the results based on the user-based quality metric in a user study?
- **b.)** Is the measured effect perceived by the professional users when carrying out their task?

3. METHODOLOGY

In order to answer our research questions, we need understand the behavior of crowd workers when they are confronted with a simplified expert task (RQ 1). The criteria for the chosen task are the following: the task has to be a recognized expert task and expert data must be available to compare the user judgements to (RQ 1a.). Additionally, the simplification has to be feasible in an automated way, e.g. selecting potential correct answers with machine learning.

By showing items repeatedly, we are able to measure changes in the users' performance over time (RQ 1b.). Our approach is to present a game where the users have to select the correct answer from a set of five candidates. This allows us to give feedback about the correctness of the choice as points. For the experiment, we selected the candidates manually but in a way that is similar to what we can expect from automatic classifiers (RQ 1c.). One condition simulates a hypothetical perfect classifier (P@5 =100%) to create a baseline to compare against. A second condition simulates a realistically performing classifier (P@5 =75%) by removing the correct candidate in one out of four cases.

Therefore, we choose a specific user group (humanities researchers) and a specific (research) task and investigate which quality issues have the strongest effect on their work. To understand how they conduct research, we conducted an interview with a cultural historian. The insights gained will be used to develop questions for semi-structured interviews with further researchers from humanities research institutes (RQ 2a.). The aim of this interview study is to create a list of research tasks that are strongly influenced by quality issues of data collections (RQ 2b.). The researchers are also interviewed on the importance of the specific tasks in their research process (RQ 2c.). These pieces of information enable us to rank the research tasks mentioned by the interviewees according to how important it is to improve them.

Next, we look into the question how we can crowdsource the improvements for data issues on a specific task. From our list of research tasks and quality issues, we choose one quality issue to improve upon. We design a crowdsourcing task that extends or corrects an existing data set that is known to be problematic for a selected research task (RQ 3a.). We aim at obtaining a large quantity of annotations that contribute substantially to reduce the data quality issue. We analyze the quality of the contributions and measure the gain in quality with our user-based quality metric (RQ 3b.). From the experiences gained in the crowdsourcing process, we will gain insights in the tradeoffs we have to deal with between the complexity of improvements we aim at and the feasibility to crowdsource them.

We will conduct a user study with professional users who are asked to perform the chosen research task on the original (baseline) and on the enhanced data set. By observing their behavior and with a questionnaire we aim to find out how useful they perceive the crowdsourced enhancements of the data (RQ 4a. and RQ 4b.) and whether this is also reflected in the user-based quality metric.

4. PROGRESS MADE SO FAR

The Rijksmuseum Amsterdam is interested in extending the available data on their collection items by crowdsourcing precise descriptions. For our first study, we chose the annotation of subject types (such as landscape, history painting or still life), a task which is usually performed by museum experts. A study conducted by [9] showed that the classification cannot be successfully done by automatic classifiers. They can, however, provide a set of candidates that is likely to contain the correct class.

We investigated whether crowd workers can perform a simplified version of an expert task if they are given assistance and how well they perform compared to experts [8]. We showed that the crowd workers' contributions were largely in line with the experts' judgements and that some cases of strong disagreement indicated need for re-evaluation on the experts' side.

To make the task feasible for crowd workers we reduced the complexity (limited the set of candidates they choose from) of the task and provided feedback to the users. This feedback is based on annotations made by experts from Rijksmuseum Amsterdam on images taken from a dataset created by the Steve Tagger Project[1]. It proved to help the users to improve their performance.

The analysis of the obtained data shows that users improved during gameplay, but that they need to be trained on a data set with expert feedback which allows to always present the correct candidate. Aggregating the user judgements largely removed deviations from the experts' judgements. Persisting disagreement indicated need for metadata on the users' side or incomplete / incorrect judgements by the experts. We therefore suggest to feed cases of strong disagreement back to experts for re-evaluation.

5. FUTURE PLAN

We are currently interviewing (e-humanities) researchers who use large digital collections, such as the newspaper archive[2] of the National Library of the Netherlands. By closely looking at the different phases in their research and the extent to which they use digital sources, we will gain insights into their perception of how useful digital sources are for them. We are particularly interested in finding out how error-prone sources of data (be it automatic processing of documents or data collected through crowdsourcing) influence their work. We will investigate the different types of tasks they perform during their work and from that we will identify tasks that are to a considerable extent affected by data quality issues (RQ 2a. & 2b.). As a result, we expect to identify a number of data quality issues that we can prioritize according to their usefulness and importance (RQ 2c.) for professional users and feasibility to crowdsourcing improvements. To measure the quality of the original data and the potential improvements, we develop a quality metric based on the needs of the professional users (RQ 2d.). This will be completed by the end of 2014.

By the beginning of 2015 we expect to start designing a suitable crowdsourcing task (RQ 3a.) to gather the improvements for the chosen data quality issue and conduct the data collection. We use the previously defined user-based quality metric to judge the gain in data quality (RQ 3b.). The crowdsourcing task will most certainly require us to reduce the complexity of the task in order to make it feasible for experts. The insights we will gain from the design process and from the analysis of the crowd contributions will allow us to evaluate the tradeoff (RQ 3c.). This will be completed midyear 2015.

We will then validate the quality judgements we made based on our user-oriented quality metric. We will conduct a user study with professional users to find out whether the gain of quality we measured is actually perceived by them (RQ 4a. &4b.). The results are compared to a baseline user study conducted on the original error-prone data. We expect this to be finished by the end of 2015.

Acknowledgements

This publication is funded by the Dutch COMMIT/[3] program and the European LinkedTV[4] project.

6. REFERENCES

[1] D. M. Berry. The computational turn: Thinking about the digital humanities. *Culture Machine*, 12(0), 2011.

[2] M. Bron. *Exploration and Contextualization through Interaction and Concepts*. 2013.

[3] T. E. Clement. A thing not beginning and not ending: using digital tools to distant-read Gertrude Stein's The Making of Americans. *Literary and Linguistic Computing*, 23(3):361–381, 2008.

[4] C. Eickhoff. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 53–56, New York, USA, 2014. ACM.

[5] F. Galton. Vox populi. *Nature*, 75(1949):7, 1907.

[6] J.-R. Park and Y. Tosaka. Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging amp; Classification Quarterly*, 48(8):696–715, 2010.

[7] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, Apr. 2002.

[8] M. C. Traub, J. Ossenbruggen, J. He, and L. Hardman. Measuring the effectiveness of gamesourcing expert oil painting annotations. In *Advances in Information Retrieval*, ECIR 2014, pages 112–123. Springer International Publishing, 2014.

[9] S. Wouters. Semi-automatic annotation of artworks using crowdsourcing. Master's thesis, Vrije Universiteit Amsterdam, The Netherlands, 2012.

[10] H. I. Xie. Evaluation of digital libraries: Criteria and problems from users' perspectives. *Library & Information Science Research*, 28(3):433 – 452, 2006.

[11] H. I. Xie. Users' evaluation of digital libraries (dls): Their uses, their criteria, and their assessment. *Inf. Process. Manage.*, 44(3):1346–1373, May 2008.

[1] http://tagger.steve.museum

[2] http://www.delpher.nl

[3] http://www.commit-nl.nl

[4] http://www.linkedtv.eu