Quality Assurance in Document Conversion: A HIT?

Christoph Becker
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/~becker

ABSTRACT

This paper discusses challenges and opportunities of using human computation and crowdsourcing for the task of quality assurance in document conversion processes and proposes a hybrid computer-human system approach. Digital content is never presented to a user directly, but always needs an intermediate presentation that is generated through an algorithm (such as a document viewer) that interprets data. When converting data such as documents, the question of authenticity of the derived representation of these documents requires a comparison of the intellectually perceivable outcome of different interpretations. Such Quality Assurance is a key obstacle to scalability in document conversion processes. Currently, there is a severe lack of scalable techniques. We argue that this comparison is a Human Intelligence Task (HIT). To investigate the feasibility, potential pitfalls and key challenges in leveraging the wisdom of the crowd for this task, we have conducted several pilot experiments. We describe and discuss these experiments, and identify a number of key challenges that need to be addressed. In particular, we discuss the questions of motivation; task semantics; presentation and interaction design; and quality control. Finally, we outline a proposal to address these challenges in a hybrid computer-human system.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries

General Terms

Design, Experimentation, Human Factors, Measurement

1. INTRODUCTION

The massive move towards digitization and digital production has opened enormous possibilities to create, manage and use information. This has in particular enriched the many ways readers interact with digital documents in general and books in particular. However, digital content represented in a code always requires a certain environment to be accessed. Without this environment, valuable information turns into worthless data streams. The environments we use to access digitally encoded information change at a rapid pace. Hence, be it browser releases or office software, *longevity* is not a property commonly attributed to the digital world.

When we convert a document to render it in an environment different from the original, we create a new representation of a piece of content perceived as a semantic unit. The intellectual content is supposed to be unchanged, but how do we know it is? Both representations are interpreted using different sets of algorithms. The fundamental problem of Quality Assurance here is to judge the equivalence of information that is encoded in different representations and interpreted by different algorithms in order to present a rendering or performance.

Intellectual aspects of an original performance missing in a derivative performance will often not be discovered through superficial checks and commonly used testing methods. The complexity inherent in such interpretation is very high and the variation in input data extremely large. The sheer volume of data in current systems, the combinatorial explosion of possible input features and the variation in rendering environments together cause a serious scalability problem in digital libraries.

Document conversion processes are relevant in a number of scenarios. The original background of this work is motivated by digital preservation. The field of digital preservation is concerned with keeping digital information authentic, understandable, and usable, through time and across changing socio-technological environments. Essentially, the fundamental problem addressed is a misalignment of technology: To render information usable to any human, an algorithm needs to produce an interpretation that can be percieved by the human. Digital preservation is in this sense often seen as a case of interoperability through time.

From this perspective, Quality Assurance as an operational capability is the ability to deliver accurate measures that quantify the equivalence of performances (renderings) of content by measuring their properties and comparing them to each other to quantify their equivalence corresponding to requirements. This requires the ability to measure properties, i.e. extract relevant features from the content. Essentially, measuring these can be based on algorithmic analysis of bytestreams; perception-based analysis of renderings; and observation of the interactive behaviour of a certain object

BooksOnline'11, October 24, 2011, Glasgow, Scotland, UK.

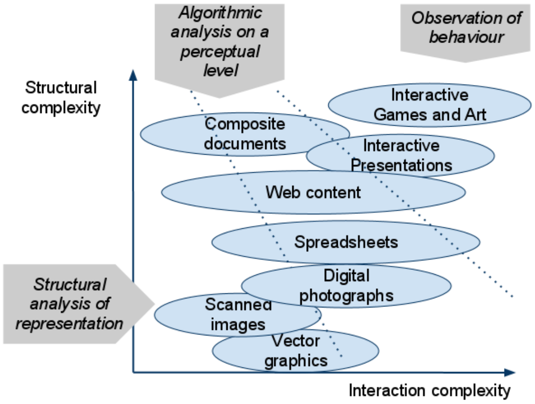

Figure 1: Approaches for Quality Assurance

in a certain environment. These complementary approaches are visualized informally in Figure 1.

We observe that the task of perceptual analysis can be split into the interpretation and rendering of encoded documents and the analysis of this interpretation. We can thus formulate the analysis as a Human Intelligence Task and address QA in a hybrid computation system[1].

We have investigated the feasibility of leveraging crowdsourcing approaches to tackle the problem. We constructed a test set of documents and conducted a series of experiments to investigate the issues and potential. We will shortly outline the experiments and draw some lessons learned. We then propose an approach for large-scale experimentation and development of an integrated QA platform. This would leverage the wisdom of both the general crowd and the specific expertise and motivation of the much narrower community of digital preservation practitioners.

The next section will outline related work in the areas of document conversion and quality assurance in the digital preservation context, and discuss its relation to existing work in crowdsourcing and human computation. Section 3 will discuss the experiments we conducted to explore key issues. Section 4 discusses lessons learned and outlines challenges to be addressed in order to make crowdsourcing a viable solution for this yet unsolved problem area. Finally, Section 5 presents an outlook on our vision to tackle the challenges described.

2. IS QUALITY ASSURANCE FOR DOCUMENT CONVERSION A HIT?

Evaluation of quality of certain representations is a key question of digital preservation research, which has produced a sustainable framework for quantitative evaluation of alternative representations and the conversion processses between them. The approach is based on a variation of utility analysis, combined with controlled experimentation and automated measurements. The key elements of the assessment are a weighted hierarchy of objectives leading into measurable criteria and a utility function for each criterion specifying the assessment for the range of possible values. These two aspects are modelled in a so-called objective tree which is constructed carefully to reflect a specific decision making context and forms the nucleus of evaluation. Decision making preferences are highly context-dependent, thus the combination of objective evidence and situational context-dependent assessment is essential. Conversion processes are evaluated against the goal hierarchy defined in this objective tree. The effects of carrying out candidate preservation actions are measured in controlled experiments using a representative set of sample objects [3].

The digital preservation community has invested considerable efforts in ‘characterization tools’, programs that perform information extraction on digital content for quality assurance purposes [4]. All these approaches employ static analysis on the representation itself, e.g. on the Word files. So far, however, these efforts have not yielded any scalable, automated and reliable quality assurance tools that could truly assist even the case of an electronic book, much less a more complicated one such as compound objects. A major obstacle is that a large part of the semantics is often contained not in the objects themselves, but only realizes itself in a rendering environment. This makes it difficult to arrive at meaningful results by applying merely static analysis means, a symptom which increases correspondingly with the structural complexity of files [12]. The problem needs to be considered in a multi-dimensional view: Comparing the original document A in a new viewer environment to the performance in the original environment cannot be distinguished from viewing a derived document B in the original viewer. A derived document in a different viewer again represents an entirely different performance.

Most current QA techniques operate on the level of file formats, trying to interpret the extracted properties from different formats and compare them to each other. This is faced with two major challenges: (1) The mapping of properties between formats is very often not homomorphic at all; even worse, there is often no clear way of creating such a mapping at all. Consider an Office Open XML document with a table, converted into PDF. In OOXML the table is clearly identified, but a PDF extractor will have considerable difficulties in recognizing it, depending on the way the PDF conversion tool has represented the table in the document. (2) The multitude of formats and their variations makes this kind of property extraction computationally intensive and error-prone.

Considering the process, the reasons become clear: Essentially, a full analysis just means creating yet another viewer environment. Instead, it is possible to achieve better results by evaluating characteristics on the perceptual level and analyze a trusted interpretation produced by a reliable, well-tested tool on a standardized reference platform. The viability of such an approach for the case of digital photographs in raw image formats is demonstrated in [2].

It is clear that prior to exposure to such an interpretation, the properties of any digital object are entirely unknown. This causes a fundamental problem: There is no ground truth that can be used safely to evaluate approaches and system parameters on a large scale. This differentiates the design problem from scenarios with known ground truth, where key experiment parameters can be explored systematically in large-scale experiments [8].

Simply reducing the coverage of measurements to a level that is practically feasible on the technical level is not sufficient in many cases where high requirements are posed on trust and authenticity. In many scenarios, legal requirements impose strict constraints regarding authenticity and

[1] A HIt is “the task you ask a Worker to complete. It may be a task that is inherently difficult for a computer to do.”[1]

verification of processes to avoid litigation. This implies that we need a high degree of trust also in the measurements used in the course of the process. We thus investigate the potential to complement automated analysis with human computation.

Human computation is a rapidly growing field [13], and a recent contribution even mentions the usage of crowdsourcing the OCR aspect of a massive digitization project [5]. Quinn suggested classifying human computation systems according to motivation; quality control; aggregation; human skills employed; process order; and task request cardinality [13]. For our preliminary investigation, a number of aspects are particularly relevant. *Motivation* is in the simplest case based on micro-payments, but could also use enjoyment and, within a community, altruism and reputation. *Quality control* and *HIT design* as the most challenging aspects will be discussed in detail below.

3. EXPLORATORY EXPERIMENTS

From the discussion above, we can draw several observations. Comparing pages of documents is a computationally challenging task that can be broken up into smaller subtasks. For every pair of objects, the answer is in principle decidable, but the computation that is needed may be challenging. On the other hand, standardized reference renderings can produce reliable interpretations of objects that we can generate automatically and use as intermediary artifacts to support comparison.

We conducted a series of experiments using CrowdFlower[2], where pairs of pages from original and converted documents were analyzed and compared. Pages were presented for comparison by printing the documents on reference platforms to PDF and converting the resulting PDFs to images. Some documents contained tables, some were text-only, some contained diagrams and some footnotes. We used several conversion processes, each of which introduced a number of errors. The viewer environment in each case was the standard viewer for the target format. Conversion from OpenDocument to Google Docs led to shifts in paragraph positions, missing footnotes, and missing diagrams. Conversion from MS Word binary files to Office Open XML [7] led to minor changes in font sizes and shiftings in the text. Conversion from MS Word binary files to Office Open XML *and back* led to slight shifts in the positions of footnotes and paragraphs, and minor variations in font sizes. Additionally, we introduced manual errors in a subset of the documents to create conversion-like errors with specific characteristics.

Each task consisted of one page in two renderings, i.e. a pair of images, and a number of questions about the differences between the images. Corresponding to standard practice, we set the payment per task to be 10 cents.

In a first preliminary experiment run, we asked workers to compare documents for a number of properties. The main instructions were along the lines of 'help us to find differences in the following documents'. The page design presented two pages side-by-side and a series of questions asking for changes in the positioning of images and tables, differences in footnotes, and textual differences. Since a full-sized view of pages was not possible on a normal screen, links to magnify each page were included. The amount of quality control was minimal, with just a small fake captcha (We asked for the last word on the page, but had no annotations to verify it). The results were correspondingly unusable: Not only were the semantics of questions such as 'Are there textual differences?' apparently too subjective and confusing, also the layout of the question design: The majority of workers did not even look at the magnified version of the documents. Even with substantial consensus built in, results were close to random.

Even with this small experiment, it became clear that there are two major obstacles: (1) The semantics of *difference* are very vague, and (2) sophisticated interfaces are needed for document comparison. The latter, however, require screens that are bigger than those of typical workers on platforms such as AMT. In a second run, we thus dropped the goal of having workers complete the *Quality Assurance* task as a whole. Instead, we partition the task into the *analysis* task and the *comparison* task and aim for a hybrid computation system where the humans analyze documents and produce unambigous labels about the properties, which can then be easily compared in an automated way.

We conducted a small experiment using gold questions, i.e. questions with known and documented answers, and consensus. The language of the questions was chosen more carefully to be clear about properties and labels, and the questions were very simple, yet not answered by common automated analysis tools. For example, we asked about the number of footnotes on a page. Convergence was good, and results were usable. This experiment thus confirmed tentatively that simple analysis tasks are feasible. However, it did not involve sophisticated comparisons or large-scale evaluation. Obvious improvements include the following:

- Use a qualifying entry test that trains participants in the semantics of differences in a playful way and uses test data with known ground truth to verify the participants' ability to analyze documents.
- Log worker behaviour in detail to optimize the use of captchas, detect bias and cheating patterns, and discover pitfalls in HIT design.
- Given that property descriptions of objects are just a specific type of labels, algorithms for improving labeling quality such as [6] can be integrated easily.

While these will certainly increase quality for the simple analysis tasks, larger questions about the system design remain.

4. DOCUMENT CONVERSION AS A HUMAN INTELLIGENCE TASK

Comparing the document comparison task, and in particular the much more approachable *analysis* task, to other human intelligence tasks addressed through human computation systems, a number of challenges present themselves. These are very much concerned with the question of HIT design and the interplay between human and machine computation in hybrid systems.

- **Semantics and ambivalence**. The semantics of comparing documents are very ambigous, since the distance between documents and their elements needs to be judged on a variety of dimensions at the same time. Moreover, some of these are not independent: For example, a part of a paragraph may be missing that also

[2] `http://crowdflower.com/`

contains specific formatting elements. Finally, in many cases the semantics depend on the usage context and the users' knowledge.

- **Renderings**. As discussed above, the sheer size of rendering pages presents a simple, but hard problem. One way to address it may include breaking up pages into their composing elements and having these judged independently, similar to previous approaches to analyzing forms [11]. This would reduce the size of the problem to more manageable and verifiable pieces, and would also ease the provision of gold answers for quality control.
- **Multiple pages.** The problems related to segmentation and complex differences are only exacerbated when considering multi-page documents: Shifts of content across pages are far from uncommon and may cause reports of massive changes, even though only one empty page has been added (as happens frequently). While this appears as a daunting task, on the other hand it is exactly an argument for using a hybrid human-machine computation approach: Spotting shifts across pages is a prime example of a task that will for a while be much easier to complete for humans. However, this requires a more sophisticated interaction design than simply presenting page-oriented views and separating pages.
- **Motivation.** Using mechanized work has a low-entry barrier and may be 'good enough' for many purposes. However, community-based engagement has been shown to yield substantial benefits for well-defined tasks in domains such as online books [9] and may be much more effective.

Focussing on the design of HITS, we can see a number of ways to tackle the challenges outlined:

- **Comparison aids.** Approaches for supporting users' judgement about page similarity have leveraged OCR techniques to visualize comparison aids [12]. This may introduce unwanted bias and has technical limits, but has enormous potential to increase accuracy and productivity when employed properly.
- **Controlled reformatting.** As an intermediate or supportive step towards controlling the parameters of the system until performance is stable and quality control mechanisms in place, it may be possible to contrast controlled intentional reformatting (with full control over the formatting) with less controlled conversion.
- **Task partitioning.** Address the semantic overload of a problem such as *difference between text blocks*, where the difference could be in content, appearance, structure, or even behaviour, with a further split-up of tasks into small digestible portions:
    1. Partition a page into composing elements by producing markup: Where are the paragraphs, tables, headings,...?
    2. Judge the equivalence of element pairs: Is this pair of paragraphs supposedly representing the same paragraph? (Alternatively, analyze properties of one element at a time, i.e produce labels for elements.)
    3. After the equivalence of content has been confirmed for pairs of elements: Judge the similarity of structure and appearance for each element.
    4. Finally, judge the positioning of elements: Are elements positioned correctly in the copy? This will in general require a comparison, but once the equivalence of elements has been confirmed by a human, it should be easily computable.
- **Collaborative annotation and input agreement**: Let participants create a hierarchical model of a page by composing a tree using predefined template elements. This can be done alone or in teams. A number of winning conditions come to mind: For example, one player or team could see a number of such trees constructed at the same time and pick pairs, even in the style of a Memory game[3]. Alternatively, participants could have to reach pairwise agreement on the input in the style of TagATune [10]. In this case, the input agreement would be a side effect used only for entertainment purpose, while the produced labels are the valuable output. Again, if the labels produced are rich enough, they should be sufficient input for a similarity algorithm to compute pairwise distances.

In terms of system design and motivation, three options are possible:

1. **Standard task marketplace.** A dedicated platform such as an extension of the conversion portal [12] could be used to dynamically integrate sophisticated comparison aids, and the task partitioning approach described above should be explored.

   This scenario lends itself well to an exploration of system design parameters. However, it could also be integrated in an evaluation environment commonly used by typical service requesters: For example, the planning tool Plato[4], which is used in digital preservation, could dynamically generate the job submission packages and deposit them to the marketplace to request asynchronous evaluation.

2. **Community Game.** The structure of the problem lends itself well to a *game with a purpose* based on output agreement [14]. Take a number of documents, a number of top-down defined criteria to be judged, a number of users and a number of converted representations of the original documents. Let the users evaluate the faithfulness of representations in a game with a purpose. For example, a group of explorers is searching for errors, while a group of reviewers validates found errors similar to the Book Explorer game [9]. An alternative model would be based on collaborative annotation: Have two teams find differences in documents – the one with more 'points' wins. This setting should be particularly strong in addressing semantic ambivalence.

   The long-term problem with this approach is the critical mass of a community required to play the game: It may be possible to draw this community for a short while, but no sustainable scenario comes to mind that

[3] `http://en.wikipedia.org/wiki/Memory_(game)`

[4] `http://www.ifs.tuwien.ac.at/dp/plato`

would promise continued long-term engagement. However, such a community game would have great potential if used like in the INEX example, where a community of interest created a benchmark data set [9]: Having document sets with ground truth labels would provide tremendous value to the community.

3. **System integration.** Quality assurance mechanisms in the form of small analysis micro-tasks can be integrated into existing information systems. In contrast to the community game, the benefits relation for participants is clear, and the approach would engage the commmunity exactly at a point of interest. For example, simple micro-tasks could be embedded in the access interfaces for digital libraries that provide content. Answering one question per session (or one question and one captcha) may be perfectly acceptable for many users if they know they contribute to an overall improved user experience for the entire community. This appears a perfectly viable approach with a low entry barrier.

5. OUTLOOK

QA for conversion processes requires us to validate perceptual-level representations of documents. The combination of algorithmic analysis and human computation in a hybrid approach clearly has substantial benefits to offer. However, there is a number of dimensions along which we have to position our approach. The question is, where is the sweet spot?

We believe that integration of crowd-sourcing functionality into digital library systems is a very promising scenario and thus propose to pursue this path. Additionally, we suggest to involve the digital preservation community in a large-scale experiment based on the conversion portal [12]. The design phase of this experiment can rely on additional evaluation experiments on a standard marketplace. The results would have the potential to fill a gaping hole in digital preservation research and practice: The fundamental absence of benchmark data sets is severely limiting progress in QA methods [3]. These solid large-scale benchmark corpora can then be used to systematically verify or even train algorithms. The proposed system design not only supports conversion processes in digital library systems and preservation scenarios, but should also further insights into cross-device reformatting scenarios.

The evaluation framework described in [3] separates objective evidence (such as the exact quantified difference measures we are discussing) from subjective assessment of their utility and thus enables decision makers to model their exact preferences. These decision makers need to represent the stakeholders involved. However, in the context of large digital library systems such as Google Books, the ideal approach would directly use the aggregated preference structure of the actual users. We can explore the correlation of measures and judgements to derive relationships and validate criteria hierarchies. Take a number of measurable criteria organized in a hierarchical structure (e.g. a tree for measuring the 'truthfulness of layout to original' when converting a document) and subjective human judgement ('layout has been truthfully preserved'), i.e. the overall utility perceived by humans. Let users give scores to the representations on various levels of their goal hierarchy (from the overall score to scores for layout and interaction elements). This results not only in baseline judgement data to complement the objective ground truth of generated data sets, but also allows us to analyze the correlation of the structured hierarchy of criteria, calculated by their aggregated weighted utility functions, to human judgement. This can be leveraged to produce accurate criteria templates by relating the measurable factors that contribute to the perceived scores in a meaningful and solidly validated way.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

6. REFERENCES

[1] Amazon Web Services LLC. Requester best practices guide, June 2011.

[2] S. Bauer and C. Becker. Automated preservation: The case of digital raw photographs. In *International Conference on Asia-Pacific Digital Libraries (ICADL'11)*, October 2011.

[3] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *JASIST*, 62(6):1009–1028, June 2011.

[4] C. Becker, A. Rauber, V. Heydegger, J. Schnasse, and M. Thaller. Systematic characterisation of objects in digital preservation: The extensible characterisation languages. *JUCS*, 14(18):2936–2952, 2008.

[5] K.-T. Chen. Human computation: Experience and thoughts. In *CHI 2011 Workshop on Crowdsourcing and Human Computation*, 2011.

[6] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proc. KDD-HCOMP'10*, Washington DC, USA, 2010.

[7] ISO/IEC. *Information technology – Document description and processing languages – Office Open XML File Formats – Part 1: Fundamentals and Markup Language Reference (ISO/IEC 29500-1:2008)*.

[8] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *SIGIR'11*, July 24-28 2011.

[9] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *SIGIR'09*, 2009.

[10] E. Law and L. von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *CHI 2009*, 2009.

[11] G. Little and Y.-A. Sun. Human OCR: Insights from a complex human computation process. In *CHI 2011 Workshop on Crowdsourcing and Human Computation*, 2011.

[12] N. Milic-Frayling. Digital object characterization: Document conversion and quality assurance. In *Automation in Digital Preservation: Dagstuhl Seminar Proceedings 10291*, Germany, 2010. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik. `http://drops.dagstuhl.de/opus/volltexte/2010/2901`.

[13] A. J. Quinn and B. B. Bederson. Human computation: a survey and taxonomy of a growing field. In *CHI 2011*, 2011.

[14] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51:58–67, August 2008.