Digital Scholarship in the Humanities Advance Access published December 15, 2015

# Navigating the Storm: IMPACT, eMOP, and Agile Steering Standards

**Laura C. Mandell**
Initiative for Digital Humanities, Media, and Culture Texas A&M University

**Clemens Neudecker**
Staatsbibliothek zu Berlin - Preussischer Kulturbesitz

**Apostolos Antonacopoulos**
Pattern Recognition and Image Analysis (PRImA) Lab, University of Salford

**Elizabeth Grumbach**
Initiative for Digital Humanities, Media, and Culture, Texas A&M University

**Loretta Auvil**
Illinois Informatics Institute, University of Illinois

**Matthew J. Christy**
Initiative for Digital Humanities, Media, and Culture, Texas A&M University

**Jacob A. Heil**
Mellon Digital Scholar for the Five Colleges of Ohio

**Todd Samuelson**
Cushing Memorial Library, Texas A&M University

**Correspondence:** Laura C. Mandell, IDHMC, 4227 TAMU, Texas A&M University, College Station, TX 77843-4227, USA.
**E-mail:** mandell@tamu.edu

## Abstract

This article discusses two major initiatives tasked with developing tools to improve optical character recognition (OCR) or the mechanical keying of texts that are digitally available only as page images. The two initiatives are the IMProving ACcess to Text Project in Europe and the Early Modern OCR Project in the USA. Because of dealing with a multilayered problem like OCR technologies and having to collaborate with radically interdisciplinary and international team members, the two projects developed techniques that we call Agile Project Management, outlined in this essay with rationales for their use.



doi:10.1093/llc/fqv062

In this short article, we discuss the challenges in project management that were confronted by two very large historical optical character recognition (OCR) projects: the IMProving ACcess to Text (IMPACT) project, spearheaded by the National Library of the Netherlands, and the Early Modern OCR Project (eMOP), run by Texas A&M University. The goal for both projects was to solve the OCR problem for early modern and European historical texts printed in fonts that are difficult to read for humans, let alone machines (e.g. Blackletter). Until print become modern around 1820, letters varied in their placements on lines and pages were plagued by bleedthrough as well as inadequate impressions; even after 1820, a multiplicity of typefaces makes it difficult for machines to accurately transcribe text from page images into keyed binary characters. In both cases, the work plans we set up for each grant-funded project, the milestones and deadlines that we promised our collaborators and funders to meet, depended upon our belief that we would systematically eliminate OCR problems by fixing images, training and improving OCR engines so that they could read anything, and developing post-processing routines that would augment these early results. Also in both cases, we discovered that neither OCR nor collaborators work that way: we were forced into acquiring considerable agility in project management techniques, revising goals and systems for working with international teams of collaborators as we went along.

In a seminal article about 'the Importance of Failure', John Unsworth has said, 'If an electronic scholarly project can't fail and doesn't produce new ignorance, then it isn't worth a damn' (Unsworth 1997). One might say of the IMPACT and eMOP projects that they 'failed' to solve the OCR problem, and therefore are 'worth a damn'. For example, one of eMOP's successes is that its OCR transcriptions achieved a level of correctness close to the most expensive commercial OCR vendors. In contrast, one of its 'failures' is that documents published in the 16th and early 17th centuries are only 68% correct when compared with ground truth. However, that figure has been lowered by pages that are completely unreadable by the OCR engine—the images are black, or the imaged pages were blotched, torn, or folded—so that eMOP developed an algorithm for determining which pages are the culprits. One of eMOP's final results is thus a database listing documents that need to be reimaged. In this case, failure to achieve a high level of correctness has produced new knowledge about what documents are not in fact, as libraries may falsely presume, digitally preserved.

However, one might say that these two initiatives, eMOP and IMPACT, have not failed at all, that agile project management involves adjusting goals. Instead of developing one or two OCR engines that could read anything, we have had to develop complex processes for OCR management, both of which include producing multiple systems for improving OCR results beyond what scholars and indeed companies have been able to achieve so far. While goals need adjusting based on discoveries made about technical possibilities, they also need to be adjusted based on social possibilities, based on the fact that experts who work together on international, multiinstitutional, interdisciplinary projects face enormous challenges in getting their various knowledges integrated into one technological process. Such was the case with IMPACT.

The IMPACT project[1] was a large-scale European research project undertaken from 2008 to 2012 with the aim to significantly improve the state-of-the-art of historical document recognition by pushing innovation in OCR and language technology for historical document processing and retrieval, and by sharing expertise to build capacity in text digitization. The IMPACT consortium comprised twenty-two partners from Europe, Russia, and Israel, including libraries, research institutes, and commercial companies. Due to the large scale and running time, and the various parallel activity streams in the project, the project management was following a strict waterfall planning and was organized according to the principles of the PRINCE2[2] methodology.

One of the main objectives in the original project plan was to develop a highly complex integrated and adaptive OCR software system that would allow collection holders to recognize historical documents of various kinds. However, following research undertaken in the early stages of the project, it became

apparent that no single system could ever deal with the wide variety of challenges and the complexity of historical documents. For example, novel approaches for layout analysis and segmentation that were developed for old books do not apply equally well to historical newspapers. Similarly, different solutions had to be followed for historical spelling variation in different languages: while for English a large and sufficiently rich database of historical spelling was available in the Oxford English Dictionary, the same was not true for German. Moreover, historical German features many more spelling variants and greater inconsistency in texts that are only a century old, thus requiring a more flexible implementation of dictionary corrections that takes into account document subject as well as period specificity. Such and similar granular challenges in the technical approach were discovered only as part of experimentation with the document and technical resources developed in the project ramp-up phase and subsequently required changes to the work plan and schedule for deliveries across several areas of the project.

In addition, the project staff comprised experts from various domains such as digital library developers, pattern recognition researchers, and computational linguists, all of whom had very different approaches to solving the main problems in their own field. For instance, the document image analysis/recognition researchers expected to be given by the library partners, at the project kick-off meeting, a representative set of documents from their collections so that they could systematically and comprehensively identify issues affecting OCR performance and their significance as well as the distribution of various problems across the collections. This selection of representative data sets turned out to be a nontrivial task, taking almost 2 years, as the library holdings are truly vast and varied, and there are multiple institutional procedures to follow in order to generate samples from various library holdings. As a result, some major changes had to be applied in project management to guarantee a successful collaboration across sectors, and to develop technological solutions suitable to the broad range of challenges.

Two approaches were adopted to facilitate the collaboration and software development across the various groups. First, in order to make sure everyone was on the same page and speaking the same language, some measures were taken to increase the amount of sharing of expertise and ease the communication and understanding among all the groups. A buddy program was implemented where typically a software developer would be teamed up with a librarian or linguist. The buddies would have frequent communications and report on each other's work progress or results at workshops and project meetings. Social games such as quizzes about the partners were run, so as to better understand the environment and ways of thinking of different communities. This all lead to a high level of engagement between partners coming from different backgrounds and established much better lines of communication and an overall much broader experience for everyone.

Secondly, in the technical work packages, a SCRUM-based[3] agile development was adopted, and the main output was changed from an integrated monolithic system to an interoperable and modular suite of loosely coupled web-based tools ('web services'). These web-based tools then only had to follow a common specification as to how they would advertise the functions that they offered as well as the input and output formats that were supported.

This revised approach entailed a number of benefits compared to the original plan:

1. Developers of individual components could focus their efforts more on the optimization of their own component and worry less about the integration of it into design of the overall system.
2. Delays or issues encountered in the development of individual components could be de-coupled from the development of other components, thus reducing the impact of such changes to an absolute minimum in terms of the development of the overall framework.
3. Where individual components were not performing as well as anticipated or were not applicable to the wide array of documents, they could be easily replaced by others that were either readily available or more suitable to the given task.

Furthermore, to keep close track of the progress of individual components and the way they interoperate, developers across the various groups had online meetings and sprints lasting one week, and the users and stakeholders of the content holding institutions also participated in the developer meetings and technical workshops at all times. This led to a highly collaborative and experimental workflow development approach (Neudecker *et al.*, 2011) which had benefits in areas such as knowledge sharing, transparency and flexibility. The development of the highly modular and interoperable software framework also encouraged the inclusion of third-party open-source technologies, the possibility to build and adapt more complex tool chains from a wider variety of software technologies, and more suitable for highly specific requirements of particular challenges found in the large variety of historical documents. Finally, the evaluation of individual approaches and technological solutions benefited greatly from this transparent system, as individual tools could be compared with other available technical resources on a much more granular basis. Last but not least, the modularity of components has also proven to ease the difficulty experienced by other projects and third parties in the uptake of individual components, as flexible licensing agreements could more easily be implemented on the granular level of individual components than it would have been for a completely integrated system. Moreover, the technical framework that was used to tie the individual components together has proven to be of value in itself and, being domain independent, has since been released as open source and adopted by other projects such as SCAPE (www.scape-project.eu), a large-scale European Union-funded project in the area of long-term preservation.

In terms of project deliverables, having spent so much effort in collecting and labeling large sets of documents representative of library holdings and digitization priorities, and understanding the multitude of issues affecting OCR, it was clear that a major achievement was the IMPACT data set of historical images (Papadopoulos *et al.*, 2013). It has been made publicly available, through the IMPACT Centre of Competence in Digitisation[4], a unique reference resource that enables OCR researchers, developers, and content-holding institutions to form a solid understanding of the issues related to different types of documents in a given collection and to focus on solving them.

Critical to the processing and analysis of the several thousands of document pages in the IMPACT data set was the new *Aletheia* software tool (Clausner *et al.*, 2011). *Aletheia* started from the idea of creating a semi-automated layout and text correction tool and was developed into a fully functional, complete document-analysis and recognition toolkit, now used by several groups and commercial organizations. *Aletheia* enables the complete analysis of the image content of a document page, including pixel-based enhancement of a region (e.g. paragraph, word, glyph), and manual entry or automated recognition (via Tesseract) of the textual content. It also allows the annotation of any entity on the page and its detailed description. Following the completion of the IMPACT project, *Aletheia* has been in continuous development and has been used by eMOP, as will be fully described below.

eMOP ran from 2012 to 2015. In Fall 2012, Texas A&M University received a $734,000 grant from the Andrew W. Mellon Foundation for eMOP[5]. eMOP's objective was to make machine readable, or improve the readability for, 45 million pages of text from two major proprietary databases: Eighteenth Century Collections Online and Early English Books Online. Generally, eMOP intends to improve the visibility of early modern texts by making their contents fully searchable. The current paradigm of searching special collections for early modern materials by either metadata alone or 'dirty' OCR is inefficient for scholarly research (Mandell 2013). In the grant document, we described eMOP's main deliverables:

- We intend to publish an open-source OCR workflow at grant end. This workflow will contain access to an early modern font database, customization guidelines for the Tesseract OCR engine, post-processing and diagnostic algorithms, and crowdsourcing and 'scholar-sourcing' (to use Brian Geiger's phrase) correction tools.

- But the overarching goal of eMOP, a project that blends book history (Heil and Samuelson, 2013), digital humanities, textual analysis and machine learning, is ultimately to foster a community of scholars and institutions interested in the digital preservation of, and access to, these texts.

To this end, eMOP assembled an international team of collaborators from multiple disciplines.

However, eMOP too has faced problems in the implementation of our goals and processes. During Year 1, the eMOP team and collaborators quickly realized that the grant document excellently outlined milestones and goals but did not provide the level of granularity needed to complete each. For example, we argued that we would train our OCR engine in one particular font, run the trained engine on documents that were printed using that font, and prove that training helps, all in the first month of the grant as a proof of concept for our continued activities. It took us 8 months to get the data interoperable; we only finished by the end of the grant period parsing data about printers in inadequate metadata in order to be able to tell what documents were printed by whom, and thus we had to set on the backburner doing research into which fonts were used by which printers. Not only that, training our OCR engine Tesseract turned out to be no small affair. While we had believed, as had the IMPACT group, that the more examples you give an OCR engine of a particular instance of a letter in a specific font, the better, it turned out that 'more' made Tesseract perform 'worse'. It turns out that we had to find perfect instances of our letters and then type out faux documents with these perfect instances in order to train Tesseract, a feat that required our graduate-student collaborator Bryan Tarpley to build a new tool for training Tesseract, a tool he called Franken+ because it cobbles together faux documents out of many document parts (Torabi *et al.*, 2013).

To explain in detail, eMOP discovered the procedure necessary for training the open-source Tesseract engine. First, we used *Aletheia* to label and extract glyphs for OCR training. We imported those labeled glyphs into Franken+ and then selected only the best instances among them. Franken+ also allows improving those best samples in Adobe Photoshop. Once the best instances of a font are available, Franken+ types a text using those glyphs, and that text is given to Tesseract as a document on which to train, to create a training set. All these procedures are detailed on another deliverable, the eMOP web site that offers all software and training sets on GitHub as well as instructions in using Aletheia in conjunction with Franken+ to create new, font-specific training sets[6]. Creating Franken+ and further development of *Aletheia* were two of many additional and unexpected deliverables that the eMOP team built to approach its overall goal of improving the state of early modern OCR technologies.

One major success, we believe, is that the eMOP team was able to achieve a high level of correctness for OCR'ing 18th-century texts (86%) using an open-access OCR engine along with the software and post-processing routines that eMOP developed and made available to all, close to the same correctness level as has been achieved by companies that charge a great deal because they use multiple, commercial engines and the labor of human correction (89%, using the same measures as we used to calibrate eMOP's correctness).

Agility was key to this success: putting font identification on the back burner, and revamping our font training process when we realized that it would not work. The eMOP team realized that progress is continually changing in the field of OCR, and that, if big Digital Humanities (DH) projects do not adjust accordingly, we will not be able to build on each other's work. Active outreach and collaboration with institutions outside the initial grant collaborators proved crucial. We have learned that

1. challenges and failures should be consistently communicated to every individual on the team, as
2. analysis and new directions can come from team members, or the scholarly community with which they regularly engage,
3. and DH projects, even projects that rely on limited funding or grant deadlines, should allow for the discussion of new possibilities and research questions in the face of roadblocks.

To summarize what both the IMPACT and eMOP projects have learned: agile project management may involve agile development techniques, but 'releasing early and often' (Scheinfeldt 2010; see also Beck n.d.; Martin 2003) does not work well with multilayered problems involving multidisciplinary collaborations: for that agility requires inventing collaborative techniques, such as the IMPACT Buddy system, and being willing to realign milestones and goals to accommodate both social and technical problems.

## References

**Beck, K.** (n.d.). Manifesto for Agile Software Development. Agile Alliance. http://agilemanifesto.org (accessed 30 October 2013).

**Clausner, C., Pletschacher, S., and Antonacopoulos, A.** (2011). Aletheia—an advanced document layout and text ground-truthing system for production environments, In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011)*, Beijing, China, September 2011, pp. 48–52.

**Heil, J. and Samuelson, T.** (2013). Book history in the early modern OCR project, or, bringing balance to the force. *Journal for Early Modern Cultural Studies*, 13(4): 90–103.

**Mandell, L.** (2013). Digitizing the archive: the necessity of an 'early modern' period. *Journal for Early Modern Cultural Studies*, 13(2): 83–92.

**Martin, R. C.** (2003). *Agile Software Development: Principles, Patterns, and Practices*. Saddle River, NJ: Prentice Hall.

**Neudecker, C., Schlarb, S., Dogan, Z. M., Missier, P., Sufi, S., Williams, A., and Wolstencroft, K.** (2011). An experimental workflow development platform for historical document digitization and analysis. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP '11)*. New York, NY: ACM. DOI=10.1145/2037342.2037370. http://doi.acm.org/10.1145/2037342.2037370.

**Papadopoulos, C., Pletschacher, S., Clausner, C., and Antonacopoulos, A.** (2013), The IMPACT dataset of historical document images. In *Proceedings of the 2013 Workshop on Historical Document Imaging and Processing (HIP2013)*, Washington, DC, USA, August 2013, pp. 123–30.

**Scheinfeldt, T.** (2010). Stuff digital humanists like: defining digital humanities by its values. *Found History*. http://www.foundhistory.org/2010/12/02/stuff-digital-humanists-like (accessed 30 October 2013).

**Torabi, K., Durgan, J., and Tarpley, B.** (2013). Early modern OCR project (eMOP) at Texas A&M University: using Aletheia to train Tesseract. In *Proceedings of the ACM Document Engineering Conference (DocEng '13)*. New York, NY: ACM. http://dx.doi.org/10.1145/2494266.2494304

**Unsworth, J.** (1997). Documenting the reinvention of text: the importance of failure. *Journal of Electronic Publishing*, 3(2). Lessons Learned in Electronic Publishing. DOI: http://dx.doi.org/10.3998/3336451.0003.201

## Notes

1 http://www.impact-project.eu/
2 http://www.prince-officialsite.com/
3 http://en.wikipedia.org/wiki/Scrum_%28software_development%29
4 http://www.digitisation.eu
5 http://emop.tamu.edu
6 At http://emop.tamu.edu, please see the 'GitHub' and 'Software' tabs.