Crowdsourcing facing Cultural heritage of printed texts: the platform Correct (Co-operative text correction and enrichment)

Isabelle Josse ; Bibliothèque nationale de France ; Paris, France

Abstract

The platform CORRECT was created within the frame of a research and development program which gathers 9 partners coming from institutional, industrial and educational background.

In such a context, this program focuses on a wide range of technical challenges from the correction of OCR outputs to their improvements through a crowdsourcing approach. The project follows 3 main targets: first to promote a crowdsourcing approach, to fulfill the need for e-accessibility /digital access (for all users and for all access devices), lastly to develop innovative tools and approaches based on man-machine relationship.

One of the main technical challenges was to develop tools dedicated to the production of digital transcription consistent with the original documents. The next issue at stake was the possibility led to the improvement of the digital transcriptions in order to produce new versions of these transcriptions under several formats (such as e-books format or disabled people editions dedicated to disabled people).

One of the main challenges was also to recruit, mobilize and manage a great number of contributors. Reflections on such issues and challenges had been undertaken at different levels: up to what extent and how users should be involved within the project and through what kind of collaborative mechanisms; the technical requirements of platform tools (IHM and assistance/help features) and, lastly, thoughts about contents made available.

CORRECT (Co-operative text correction and enrichment)

Since January 2012, the National Library of France (BnF) is involved in the research project FUI12 Ozalid. The aim of this project is to provide a collaborative platform for digitized documents' correction and improvement. Indeed, the optical character recognition (OCR) process is not fully reliable to get the document processed consistent with the original. To achieve such a result, the only method remains to manually correct the digital document to get rid of the remaining mistakes.

The project involves a consortium of 9 partners (Orange, Jamespot, Urbilog, I2S, ISEP, INSA Lyon, University Lyon 1, University Paris 8). Orange is in charge of the project while the National Library of France manages the experiment. The National Library of France has provided documents from its digital library Gallica. Morevover, BnF asked for its users to get involved in the experiment by testing CORRECT platform features.

Technical and scientific challenges: cultural heritage printed documents digitization state of the art

A large number of digitized documents but a heterogeneous OCR quality

Several reasons can explain this diversity: expectation in OCR quality rates depends on the digitization program which differs from a program to another; the period of time when the digitization has been processed or the original document characteristics.

According to the method used, the calculation of the OCR results can vary and corresponds more to an estimation which does not necessarily take into account the whole document.

In order to obtain a version consistent with the original document, the first step implies to get a reliable evaluation of the OCR quality process. The improvement of the OCR results is mandatory if we intend to enhance the documents indexation. In the digital library Gallica, the text file of a digitized document is shown only if the recognition rate is > 60 % which is a quite low rate.

New fields regarding the digital document use

The growing asks for full text functionalities led us to find ways of improving OCR results.

Furthermore, the spreading of the mobile devices (Smartphones, tablets or digital reading) required to take into account both the diversity of the user's practices and the access modes, more specifically by taking into account a use through touch interfaces. Another requirement of the project was to develop quality criteria allowing accessibility for every kind of users (such as people with visual handicap or any other handicap for whom the offer of adapted digital books is very low [1]) and for every kind of devices.

The previous objective implies the development of both correction and enrichment tools of correction and enrichment allowing the production of digital books (under ePub format) and/or the production of digital adapted publishing (for instance under DAISY format).

3 stages of development dedicated to an ambitious and coherent project

The following objectives of indexation, distribution and accessibility of digitized documents implied different stages regarding the design and the development revealed the necessity to rhythm the development of the platform. The project was consequently divided into 3 stages:

- The first stage was dedicated to the TEXT CORRECTION and implied the set-up of an interface for text correction consistent with the original picture;
  The second stage was dedicated to the STRUCTURE CORRECTION which implied the creation of functionalities allowing the reconstruction, for a given text, of its page setting and logical reading order;
- The last stage is dedicated to the DOCUMENT ENRICHMENT with the setting up of functionalities allowing text editorial (such as indexing, sound recorded reading or annotations).

Today the project has fulfilled the second stage with the setting up of text correction functionalities.

How to consider crowdsourcing for libraries?

Setting up collaborative projects remains of innovation for libraries, even if, today, interaction is the norm of web practices [2]. Thus, while Anglo-Saxon libraries and archives services implement important projects, those of French libraries are still in an experimental phase [3].

To ask the user to correct or enrich documents is a tricky task for any cultural institution; the approach can be misinterpreted or the target audience is not reached.

Participating in this research project allows the National Library of France (BnF) to explore and test potential ways for its digital library 'Gallica' and, at the same time to define Gallica users' expectations and needs in the context of cooperation issues.

Inventing new relationships with users

A library leading a crowdsourcing project must engage in new interactions with its users and convince them to participate in this project. Rose HOLLEY, with her experience in the digital library TROVE of Australian National Library said: "*technology alone is not the answer. We need to look firstly at what people want to do, then thebasics… We need to learn the art of working 'with' our users not doing things 'to' or 'for' them.*" [4].

For the BnF, OCR improvement (phase1) leads to a better indexing of the collections, thus helping users in their search for specific documents. However, this benefit is not necessarily sufficient to mobilize contributors to OCR improvement. The project's ambition goes beyond that. The final objective is to provide tools for correction and enrichment projects, which could go as far as the publication of new versions of the document:

- Improving indexing and text mode
- Displaying the document in a digital form
- Producing an version accessible to visually impaired people
- Developping a critical edition of a digital book

The idea is to propose a platform creating multi-collaborative projects where users meet drived by a common interest (research, project, book) or affinity. Given this approach, the library provides a framework accompanying the collaborative setting up of projects It will then let the contributors work independently.

How to motivate participation

One of the challenges of a crowdsourcing project is how to recruit and engage a critical mass of contributors. Thus, a contributor must be interested in participating and, if possible, find a common interest with other contributors. Thinking about the content made available is an important issue. This requires a reflection on the profiles of people who might be interested in this service.

Several factors have led to the choice of the collections to be the object of the first experiments:

- Firstly, a selection was made from the most consulted documents, identified by statistical analyses of Gallica users. For example, the theme of the Occult Sciences often appearing on the list of the most downloaded documents, found its place in the first collections of CORRECT.
- Secondly, further analyses of Gallica users' practices, this time web-related, led to the identification of further fields: Cuisine and cuisine history (considering cuisine blogs which take as a starting point the old recipes found in works available on Gallica); documents relating to the Great War, which are the object of discussion on various specialized forums, in particular on the occasion of the World War I centenary; or again, science fiction novels (giving 19th century people's projections of themselves in the future) gather "steampunk" communities very active online.
- Finally, in addition to the types of statistical analyses mentioned above, one could identify possible fields just by guessing what could gather a wide public: local or exotic tales (Korean tales, Breton tales and short stories, Danish tales, Indian fables and tales), popular novels or, again, collections on games or on travels.

The objective of these first selections is to address an heterogeneous public (academics, students, scholars, retired peopleetc.) coming from a variety of communities (genealogists, scientists, associations supporting the disabled, Universities etc.) At the same time, the public to be addressed is to be constituted by web users or at least by potential web users able to gain a benefit from the web

How to adapt to all types of contributors

On the basis of experience gained from crowdsourcing projects by cultural institutions, it can be shown that it is often just 10% of the users who carry out the majority of the work (up to 80% in certain cases Rose Holley says) [5]. Caroline Haythornthwaite establishes a distinction between "crowd" and "community", which rests on two different models of commitment in crowdsourcing projects: the anonymous, simple and occasional commitment is different from the community commitment which will discharge more complex tasks and precise objectives [6].

The motivations for these two types of contributors are different. The 'Transcribe Bentham' project succeeded in appealing to different types of motivations. *"We attracted an anonymous crowd of one-time or irregular volunteers, along with a smaller cohort of mutually supportive and loyal transcribers. We aimed to cast our net wide by opening the Transcription Desk to all, by creating as user-friendly an interface as possible, and by simplifying the transcription process as much as we could."*[7]

The platform CORRECT intends to address all types of contributors and so to be open to anonymous users, individual correctors who will produce specific contributions based on simple tasks, even micro-tasks. On the other hand, the aim is also to allow for the constitution of a hard core of committed contributors, who interact and work together in order to deal with complex or long-term tasks.

At present, we work on the basis of well-founded principles as well as of intuitions. For the next phase of the experiment, there remain quite a few questions to be addressed: What type of organisational structure is to be given to the network? How can we motivate users on the long run? Can we motivate by appealing to a bit of challenging ?

Three developed modules, three experimentation areas

The Platform CORRECT is articulated around 3 principal modules:

1. The engine of self-checking takes care of documents coming from Gallica (digital library of the BnF) and prepares them. It estimates and follows the evolution of the quality of the OCR on the whole document by a unique measurement standard. Finally it merges the corrections made by several revisors.
2. The social network submits the documents to be corrected by the platform users, who recuperate the documents to work on them within the framework of cooperative correction projects.
3. The third module gives the tools for corrections. For the moment only the editor for text correction has been developped. This tool is independent of the social network and a user can choose to correct a document without necessarily taking part in a collaborative project.

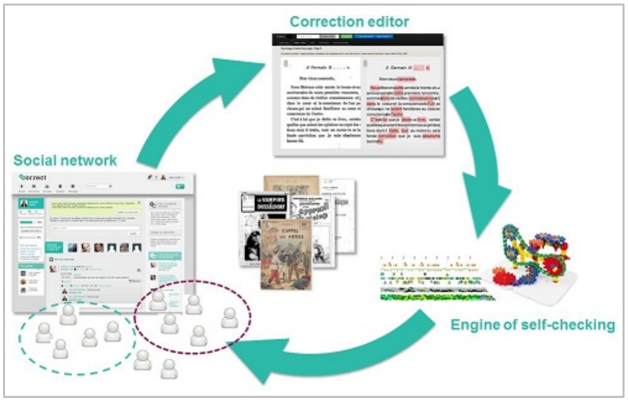


Figure 1: Diagram of the 3 modules of the platform CORRECT

This research project brings together several engineering sciences (image processing, computing, man-machine interfaces) and some human sciences (ergonomics, sociology, ecology of the social networks). It is to be considered within the context of an experimental approach to the extent that it involves, at the same time, a research activity and a series of experiments of the prototypes put in situations close to real usage.

Two experiments have already been carried out: the first experiment in April, 2013, in order to test the 1st prototype of the correction editor, the second experiment in June, 2013 further to the integration of the social network.

Development of innovating tools centered on the man-machine relation

The module of self-checking developped by Orange integrates tools which are going to help the users in carrying out their tasks:

- Automatic research for errors of transcriptions in a document supplied by OCR. This tool identifies the potential errors in a document and allows the editor to suggest to the user the corrections to make. These works were jointly led by the INSA of Lyon and by Orange.
- Assisted transcription: This tool, developped by the Orange engineers, will make it possible to extend corrections to the whole of the document.

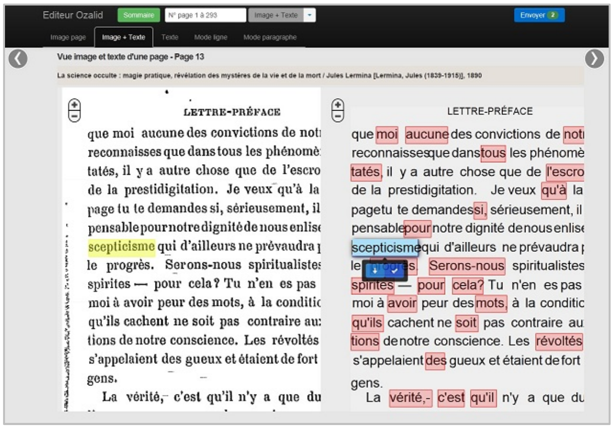
The fourth module, still under development, focuses on incremental learning techniques starting from the modeling of traces of activities. This module is designed and developped by the LIRIS laboratory of the University of Lyon 1.

These various tools are essential for the success of the project to the extent that without them the task asked to the users could be too heavy. They have to optimize the ergonomics of the platform to facilitate and simplify its use. They will be tested during the next experimentation.

Intuitive interfaces allowing various modes of appropriation

Our ambition is to offer to all these different types of public several modes of engaging with the platform. Great care has been given to the ergonomics of the correction editor, in order to offer some interfaces which are user-friendly, enjoyable and even fun to play with. Moreover the editor, developped by the company Urbilog, integrates 5 different modes of correction: Images, Images + Text, Line by line, Paragraph by paragraph, Text only.

This choice is based on the hypothesis made by researchers of Paris 8 University that the activity of correction includes different classes of situations which group together three families of activities: reading, correction and validation. The presence of several views intends to accompany these families of activities.



Vue image et texte d'une page - Page 13

La science occulte : magie pratique, révélation des mystères de la vie et de la mort / Jules Lermina [Lermina, Jules (1839-1915)], 1890

LETTRE-PRÉFACE

que moi aucune des convictions de not
reconnaisses que dans tous les phénomè
tatés, il y a autre chose que de l'escro
de la prestidigitation. Je veux qu'à la
page tu te demandes si, sérieusement, il
pensable pour notre dignité de nous enlis
scepticisme qui d'ailleurs ne prévaudra
le progrès. Serons-nous spiritualiste
spirites — pour cela? Tu n'en es pas
moi à avoir peur des mots, à la conditi
qu'ils cachent ne soit pas contraire au
tions de notre conscience. Les révoltés
s'appelaient des gueux et étaient de fort
gens.
La vérité, c'est qu'il n'y a que du

LETTRE-PRÉFACE

que moi aucune des convictions de not
reconnaissesque danstous les phénomè
tatés, il y a autre chose que de l'escro
de la prestidigitation. Je veux qu'à la
pagetu te demandessi, sérieusement, il
pensablepour notre dignité denous enlis
scepticismequi d'ailleurs ne prévaudra
le progrès, Serons-nous spiritualistes
spirites — pour cela? Tu n'en es pas
moi à avoir peur des mots, à la conditi
qu'ils cachent ne soit pas contraire au
tions denotre conscience. Les révoltés
s'appelaient des gueux et étaient de fort
gens.
La vérité,- c'est qu'il n'y a que du

Figures 2: Editor of correction integrating 5 modes of possible correction

The first experiment realized in April 2013 had for objective the observation of behaviour in a situation of correction. It was based exclusively on the use of the first prototype of the correction editor. The editor was tested by 24 people (among whom there were 2 partially-sighted people); interviews lasted one hour. For each interview, an observer followed and noted the correction activities and then carried out a post-experiment interview based on past experience.

This first experiment gave us plenty of insights:

- A good majority (87%) was positive regarding the experience carried out. This positive response relates both to the experienced interface and to the project as a whole.
- This interest is confirmed by the fact that 65 % of the users intend to be able to participate in a network of correction, even if for some of them this participation would only be occasional or would be dependent on the complexity of the corrections or on the document type.
- The presence of several modes seems appreciated in that it “makes it possible to have a choice and various points of view”, even if the Image mode and especially the Text mode were not very well perceived. However the users pointed at several anomalies of the interface (overlapping, reframing) and are still waiting for improvements on these.
- The existence of families of activities (reading, correction, validation) is confirmed by different usages according to the views. The view Lines was used for an activity of correction, while the View Images + Text was rather used for validation.

As for the interface conception, in view of captivating the users, it is clear that the interface must keep its simple, intuitive efficiency and strengthen the enjoyable and playful side of its approach.

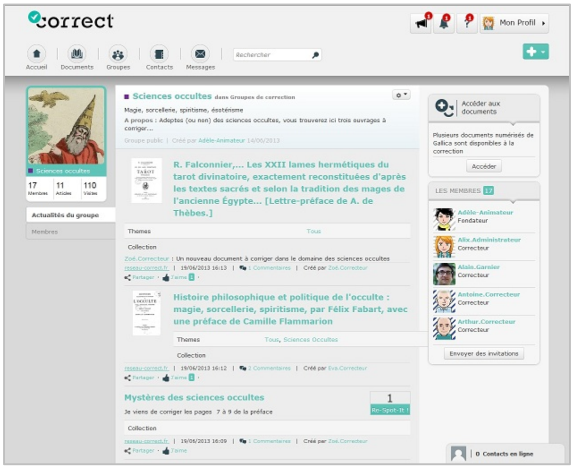
Social network as crowdsourcing lever

In order to implement these collaborative projects, the platform CORRECT depends on a social network to support and organize this collaboration. This dedicated social network is supplied by the company Jamespot.

The constitution of a network is a way to create a relationship between contributors. This relationship should be able to initiate a certain degree of emulation between users, but also favour mutual assistance and organized collaboration.

One of the primary goals is to give meaning to documents proposed on CORRECT and to the collaborative projects set up within its framework. The social network is a way to voice the various themes covered by the platform, so that the contributors can make use of them and can gather together by affinity in order to create collaborative projects.

It is also a question of allowing the contributor to engage herself in the long term. To encourage that, for the moment there is no limit to the evolution of roles within a platform. An average contributor can become an organizer of a collaborative project on which he will build a small community that he will lead according to the objectives that they will have settled.



Sciences occultes dans Groupes de correction

Magie, sorcellerie, spiritisme, ésotérisme

A propos : Adeptes (ou non) des sciences occultes, vous trouverez ici trois ouvrages à corriger...

Groupe public | Créé par Adèle-Animateur 14/06/2013

R. Falconnier,... Les XXII lames hermétiques du tarot divinatoire, exactement reconstituées d'après les textes sacrés et selon la tradition des mages de l'ancienne Égypte... [Lettre-préface de A. de Thèbes.]

Themes Tous

Collection

Zoé.Correcteur : Un nouveau document à corriger dans le domaine des sciences occultes

reseau-correct.fr | 19/06/2013 16:13 | 1 Commentaires | Créé par Zoé.Correcteur

Histoire philosophique et politique de l'occulte : magie, sorcellerie, spiritisme, par Félix Fabart, avec une préface de Camille Flammarion

Themes Tous, Sciences Occultes

Collection

reseau-correct.fr | 19/06/2013 16:12 | 2 Commentaires | Créé par Eva.Correcteur

Mystères des sciences occultes 1

Je viens de corriger les pages 7 à 9 de la préface

Collection

reseau-correct.fr | 19/06/2013 16:09 | 1 Commentaires | Créé par Zoé.Correcteur

Plusieurs documents numérisés de Gallica sont disponibles à la correction

Figure 3: Module of social network

The objective of the second experiment was the observation of behaviour in a situation of network correction.

The challenge in this experiment was to simulate the networking through a role play bringing together during two hours 10 people on site at the BnF and 22 people on remote access.

This experiment delivered several results:

- The positive opinion on the editor and the experience were confirmed
- The social network was appreciated for its user-friendliness, its possibilities of interaction and for the possibility of initiating projects.
- The collaborative dimension of the project was clearly appreciated but, at the same time, perceived as lacking sufficient visibility and clarity : “it was not easily known if there already have been corrections on the document; it is collaborative but we have not much visibility on what was made on the document “.

Conclusion and perspectives

The results of both experiments were taken into account and have led to corrections of anomalies and adaptations of the interfaces to meet expectations.

Gradual and controlled opening for a continuous experimentation

A 3rd experimentation will begin in May 2014, following the launching of the platform CORRECT inviting the users of Gallica to test it. After a few months, a user study will be carried out. Contents will be made available online only progressively and in a controlled fashion. Being able to accommodate only a restricted number of users connected at the same time (between 80 to 100 revisors), the number of subscribers will, to begin with, be limited to 1000 users

Governance and quality control

This new experiment will be the opportunity to test some of the different ways envisaged to organize the platform:

- Two main roles: corrector (isolated or in a group) / organizer (of a group). Possibility of correcting a document anonymously without passing by the social network

- Every user works on a version assigned only to him (test user). It will be merged with the corrections of the other users (of the other tests) during the generation of a new reference.
- The generation of a new reference can involve conflicts of interpretation between various users. They are managed by mechanisms of lifting of doubt redistributed among the users and which cut across the groups (ex: validation of the correction by statistics via a survey).

Concerning the quality control two think tanks are led in parallel and have to complement each other:

- The first is based on the setting up of tools for evaluation by algorithmic approach. Within the framework of a protocol of evaluation of the quality led by the ISEP Institute, an experiment is at present carried out by several partners (ISEP, I2S, Orange and BnF)
- The second approach intends to set up an evaluation of quality by the network. This reflection, carried out by Jamepsot in partnership with the BnF and the researchers of Université of Paris 8, plans to generate an index of completeness of the documents starting from the activity of correction noted on the documents. The social network will also solicit the users to regulate the conflicts of corrections.

According to the results of the experimentation and users' feedback, adjustments will be carried out following the use of the platform, which will be built together with the users and according to the use that they will make.

References

[1] Catherine Gouédard, Viviane Folcher & Nicole Lompré,"Une bibliothèque numérique à l'épreuve de la déficience : études de cas", Activités, 9(1), 78-105, (2012)

[2] Pauline Moirez, "Bibliothèques, crowdsourcing, métadonnées sociales", Bulletin des bibliothèques de France N°5, 32-39 (2013)

[3] Pauline Moirez, Jean-Philippe Moreux, Isabelle Josse, "Etat de l'art en matière de crowdsourcing dans les bibliothèques numériques", http://www.bnf.fr/documents/crowdsourcing_rapport.pdf, (2013)

[4] Rose Holley, "Crowdsourcing and social engagement in libraries: the state of play", http://eprints.rclis.org/bitstream/10760/16385/1/Crowdsourcing%20State%20of%20Play%20June%202011.pdf (2010)

[5] Rose Holley, "Crowdsourcing: How and Why Should Libraries Do it?", D-Lib Magazine. Vol. 16, n°s 3/4, (2010)

[6] Caroline Haythornthwaite, "Crowds and Communities: Light and Heavyweight Models of Peer Production", Proc. 42nd Hawaiian Conference on System Sciences. Waikola, Hawaii, IEEE Computer Society (2009)

[7] Tim Causer and Valerie Wallace, 'Building a volunteer community: results and findings from Transcribe Bentham', Digital Humanities Quarterly, vol. 6, no. 2, (2012)

Author Biography

Isabelle Josse is graduated with a DEA (Post-graduate research degree) in History and Epistemology of Economic Thought at the University Paris I Panthéon Sorbonne (1993). Since April 2012, she is Project Manager in of the Digitization service of the Preservation Services Department of the National Library of France (BnF). She coordinates the actions of the BNF in the CORRECT project.