# Crowdsourcing and digitization

Mathieu Andro

# Crowdsourcing and digitization: Presentation for the Ebooks on Demand Conference 2014 (Innsbruck University, April 11th 2014)

Mathieu Andro

INRA, DV IST, F-78026 Versailles, France. mathieu.andro@versailles.inra.fr

## Slide 1- Crowdsourcing and Digitization

Hello, my name is Mathieu Andro. I am working for the French National Institute for Agricultural Research on text mining projects and I am preparing a PhD on crowdsourcing applied to digital libraries.

## Slide 2- An opportunity

Humans spend more and more time on the Internet and they are now able to actively contribute to the development of content rather than be passive consumers.

If we consider that Wikipedia encyclopedia has received nearly one hundred million volunteer hours in a year and that the American people watch television two hundred billion hours a year, American people could annually create two thousand projects like Wikipedia rather than watching TV (Shirky, 2010).

So, crowdsourcing represents an interesting opportunity. The artwork that you can see is the isolated work of a thousand people who were unaware of the final work order.


**Ten Thousands Cents - 1000 people with no idea of the final goal**

Crowdsourcing could be defined as a type of participative online activity in which an institution or a company proposes to a group of individuals via an open call, the voluntary

undertaking of a task. The crowd participates by bringing their work, money, knowledge and experience, which always entails mutual benefit. (Estellés-Arolas, 2012)

## Slide 3- A question

Libraries have fewer and fewer resources to do the work necessary to complete their projects.

So, they could outsource to the crowd of web users, rather than outsourcing some of their tasks to providers using the workforce in low-cost countries. The online crowd includes specialists in all domains and individuals prepared to become involved. They could even fulfill objectives it would have been impossible to imagine and achieve before.

## Slide 4- Philosophy

The philosophical origin of crowdsourcing is to give humans a central place on the web as a means and a purpose. So, it may as well be regarded as being humanism or socialism. It may also be regarded as anarchism and a rejection of authority because the contribution of amateurs is equal to that of professionals and experts. Finally, it may be regarded as liberalism and its love of individual freedom, its initiative and spirit of enterprise.

This confusion in the philosophical origins of crowdsourcing is particularly evident in the field of gamification. Gamification is used to collect contributions and data users and to encourage their participation in making them play games on the web. Indeed, gamification could be a sort of Stakhanovism and socialist emulation which organized competition between factories, rewarding with medals and titles. But gamification is also like the capitalist slogan of “fun at work” and its rewards for top employees, and also its concept of “weisure”, a mixture of “work” and “leisure”.


According to some theorists, crowdsourcing allows the emergence of a contribution economy, the end of wage labor, the end of the distinction between amateurs and professionals, between leisure and work (because leisure becomes work and work a hobby), between consumption and production (because consumption itself becomes productive of value).

## Slide 5- Cons

Other scholars believe that behind the development of the ideology of Web 2.0 is hiding a new totalitarian nightmare. This movement would encourage nihilistic relativism and the denial of authority. It would lead free or underpaid labor and should be qualified as servuction because beyond any rules. For example, the Amazon Mechanical Turk Marketplace allows companies to sell microtasks, for “microsalaries” to connected workers without any legal framework and making unfair competition with traditional providers.

Others promise to tax the data in order to return to the public a portion of the value of the data they have freely produced in the form of “invisible work” for YouTube, Facebook or Google.

For professional cultural institutions, crowdsourcing could also mean the individual appropriation of the collective heritage by some Internet users to tag or give their personal and mediocre views.

You can have a look at a painting before and after restoration by an amateur.

**Ecce Homo (Elias García)**
**before and after the work of an amateur**

If a manager wants to conduct the change in a library with crowdsourcing, he will have to consider also all these arguments. However, the user who has a look at a digitized document often knows much more about the document than the librarian. He or she is, in any case, usually more qualified to do so, than the subcontractor of a low-wage country. The quality of information that the user is likely to bring is far from negligible as shown in several comparative studies.

## Slide 6- Tasks to be crowdsourced

In the field of digital libraries, almost every step of the scan chain is likely to be done by volunteers online :

- Selection of books to be digitized
- Scanning activity

- OCR correction
- Cataloging & indexing activities
- Curation
- Resource enrichment

## Slide 7- Participative digitization

Thus, the documentary selection, acquisition and digitization can be outsourced to the crowd of users:

- Internet Archive
- Commons Wikimedia
- Europeana 1914-1918

When I was in charge of digitization in Sainte-Genevieve Library, one of the two biggest academic libraries in France, we diffused about two thousand digitized books on Internet Archive, one of the three biggest digital libraries on the web.

Rather than developing a costly digital library with low visibility and unsatisfactory features, we chose to participate in a sustainable and pooled digital library, on which it is possible to download the books to EPUB and MOBI for Kindle. So you can read them on your eReaders, and not only on a computer screen. Like it was suggested by (Waibel, 2008), the old school librarians try to boost their Google ranking so people come interact with their content on their own sites. And the new school librarians will allow people to interact with their content in the places where they already work and play.

## Slide 8- Digitization on demand

Digitization funding can also be outsourced to the crowd. It is digitization on demand through crowdfunding:

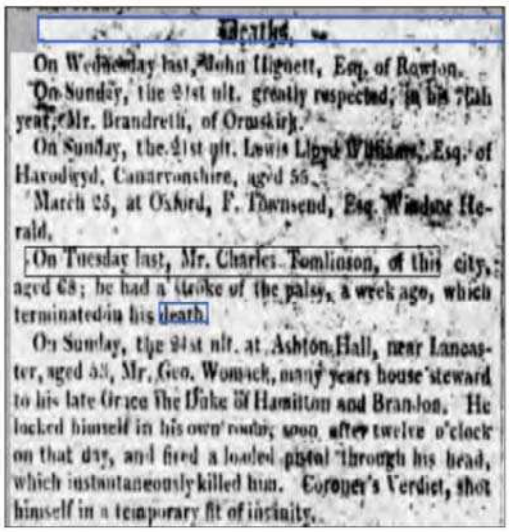
- Phénix Editions
- Chapitre.com
- Ebooks on Demand
- Adopt a book (Gallica)
- Yabé (Numalire)

## Slide 9- OCR correction

**raw OCR text**

Deaths. **Iln»rieff**, Esq. of <c .. Qn. Sunday, the till. greatly **Drandrellt**, of **Orms4\irJi**.- ~ ; ;/ ' • * On **ijfr r inn** ljjjil F iij '11 f **Havodivyd**, **Carnarvonshire**, S ; ⁴⁴" ⁴- ' « ' March **Oxford**, **F. Tfovmeud**, **Uerald.** » • V . •On Tnesdav last, Mr. **Charles. IWilinson**, this 8 ; had vf thesis#„ a week ago, which terminate<i'in his death. . / ' ▪ O'i Sunday, dJst nit. at. **AsbtCnvHall**, mar **Lancaster**, Mr.,**Geo. Worn ick**, many years house'steward hit late Once The **Hamilton** and **Brandon**. He locked himself h»oWn'r«wte<: soon. twelve o'clock" that dny, and fii»-d a loaded pistol "through Ins bead, 1 which instantaneously killed him. Coronet's Verdict, shot himself in a temporary fit of Friday week,

**newspaper image**



Deaths, &c.

On Wednesday last, John Liguett, Esq. of Rowton.

On Sunday, the 21st ult. greatly respected, in his 78th year, Mr. Brandreth, of Ormskirk.

On Sunday, the 21st ult. Lewis Lloyd Williams, Esq. of Havodivyd, Carnarvonshire, aged 55.

March 25, at Oxford, F. Townsend, Esq. Windsor Herald.

On Tuesday last, Mr. Charles Tomlinson, of this city, aged 68; he had a stroke of the palsy, a week ago, which terminated in his death.

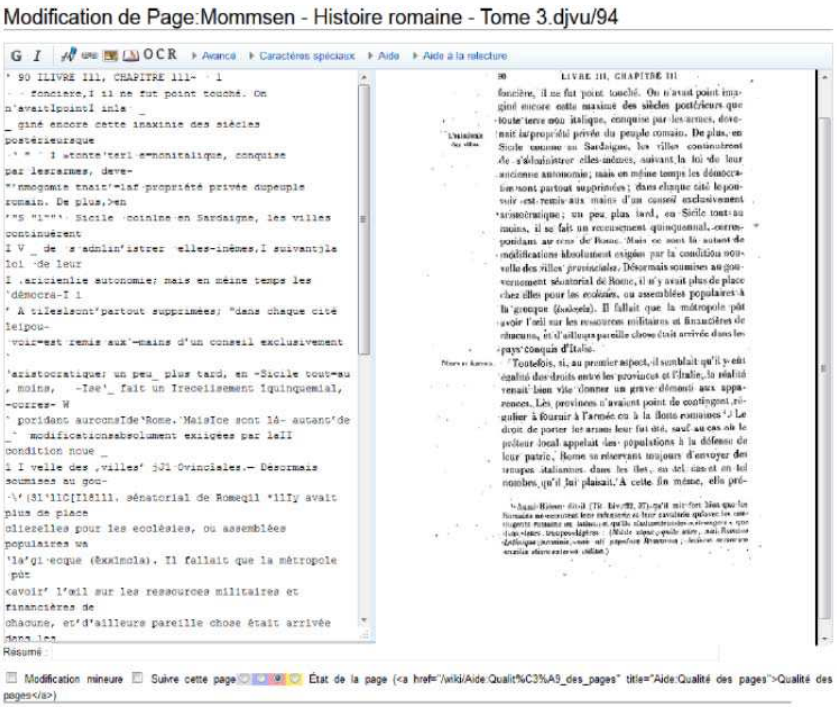
On Sunday, the 21st ult. at Ashton Hall, near Lancaster, aged 53, Mr. Geo. Womack, many years house steward to his late Grace the Duke of Hamilton and Brandon. He locked himself in his own room, soon after twelve o'clock on that day, and fired a loaded pistol through his head, which instantaneously killed him. Coroner's Verdict, shot himself in a temporary fit of insanity.

Insofar as character recognition by OCR software gives very different results depending on the fonts and the state of original documents and depending on the quality of scanning, participatory projects can, for correction, obtain high quality articles better indexed by search engines and most importantly, compatible with eReaders.

## Slide 10- Wikisource

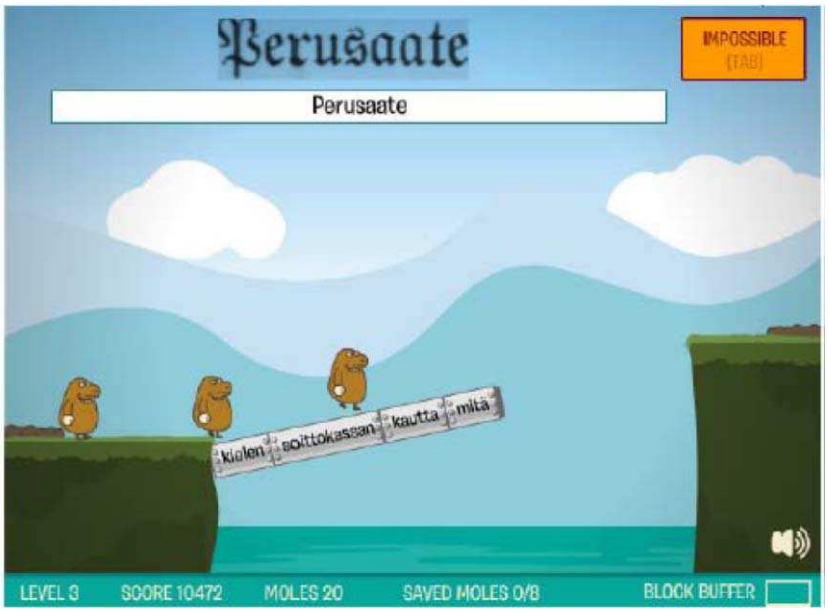
Participatory correction OCR has been experimented in too many projects to be able to evoke all of them. However, we can mention…

Wikisource. In 2008, when I was in charge of a National Veterinary School Library, I found a sponsorship with Wikipedia. One hundred thesis were digitized and diffused on Wikisource. And people have corrected texts. In 2010, 1400 books from the French National Library (BnF) were diffused on Wikisource to be corrected by the crowd.




## Slide 11- Digitalkoot

Digitalkoot involve users in the form of games on the web. In this game you have to recapture each word from the OCR to build each brick of a bridge allowing moles to across a river, avoiding a maximum of drowning and errors invariably punished by the explosion of a bridge brick.



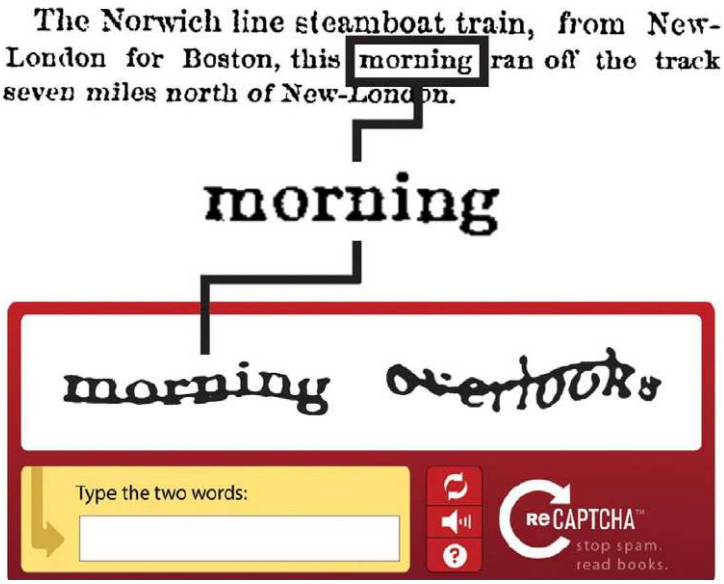

## Slide 12- Recaptcha

With Recaptcha, users must type the distorted images of two words to prove that they are not malicious robots and to create an account on a website. In doing so, they contribute to correct OCR texts from Google Books. One of the two words comes from Google Books and is not recognized by a dictionary while another word is used for safety reasons. With the

slogan “Stop span, read books”, reCaptcha allows about 200 million words to be corrected each day:

- 12,000 hours of work
- 2000 ebooks

Corrected in countries with low-cost labour, it would take about 1 million euros per day.




## Slide 13- Other OCR projects

There are many other OCR projects:

- Australian Newspapers Digitisation Program (TROVE)
- Distributed Proofreader
- Transcribe Bentham
- What’s on the menu?
- Ozalid (France)

## Slide 14- Folksonomy

Folksonomy indexing may also be requested to users as proposed, for example by steve.museum or Filckr the commons.

## Slide 15- Google Image Labeler

Google Image Labeler is a game with a purpose. With this game, you have to try to guess and find the same keyword as another partner on the web to describe and index the same image in order to score maximum points. In doing so, indexing of optimal quality is obtained.


## Slide 16- New activities for librarians

For cultural institutions that benefit from volunteer work, the development of such projects requires significant investments to communicate, recruit, manage community, motivate, reward, train, mitigate, monitor, evaluate quality and most importantly, return the data produced to digital libraries.

## Slide 17- Why users contribute?

Crowds may respond to library calls for participation. Libraries serve common goods, have a good reputation, and have a volunteer tradition. They might get work, skills, knowledge, creativity and money and so contribute to the digitization projects for reasons as diverse as personal development, entertainment, game, self-promotion or altruism.

Users can contribute for intrinsic or extrinsic reasons. Some of them want to feel useful for a group, for society, for the country, for science, for the general interest, or for a cause, or do something selfless, in a spirit of altruism and philanthropy with a sense of accountability. Others seek personal development, to grow and learn, to satisfy their thirst for knowledge. Some are looking to have fun, to play or to test an innovative approach. Others are driven by the spirit of competition, the challenge or the need to prove something. And others want to have good self-esteem, have power over things, being an author and actor or simply looking to improve their e-reputation on the web.

## Slide 18- End of crowdsourcing for OCR?

As you can see on the diagram of the Australian projet Trove, maybe a threshold has been reached and it becomes difficult to develop more crowdsourcing, especially as advances in character recognition software could progress in the future.


**TROVE (Hagon, 2013)**

And can one really talk about crowdsourcing for libraries? Contributors to cultural projects using crowdsourcing are often a defined community of volunteers whose profile is often retired genealogists for local and family history. So, it would be more appropriate to speak about communitysourcing rather than crowdsourcing.

## Slide 19- Crowdfunding experimentation at INRA

Other projects finally appeal to other types of online resources: funding resources. One speaks, in this case, of crowdfunding. From online library catalogs or from bibliographic metadata referenced on platforms, we ask the crowd, institutions or societies to fund the digitization of books via links referring to a payment interface. Documents are then digitized on site by an operator, or sent to a scanning workshop by shuttle or by mail.

At INRA, we are experimenting the Numalire Project. For my PhD, I am working with this company to know more about scanning without prior agreement with the library, scanning without a preliminary quote, and in situ scanning with mobile scanning equipment.

## Slide 20- Crowdfunding

Once scanned and put online, documents can be labeled with their sponsors and offer a return on investment in terms of advertising and web traffic. Return on investment may be particularly interesting in the case of books which can be accessed a lot of times. So, public money can focus on documents of historic or scientific interest and let private money fund digitization of books of business interest.

Libraries are now able to offer their users digital copying services, without having to bear the cost and so they are able to complete their digitization programs. They can now outsource the difficult work of identification and selection of documents which, in their documentary heritage, should be scanned and thus opened up and shared with the general public.

## Slide 21- Conclusion

Opening to amateurs can represent for libraries, as for businesses, an important source of innovations and inventions. Amateurs are not trying to reproduce the established business models with which professionals have been trained, they can bring innovative breakthroughs.

Thus, according to Von Hippel, forty six percent of U.S. companies which have survived for at least five years were created by a single user.

## Acknowledgements

Many thanks to Christine Jung for her corrections and for entertainment for the English speech.

## Bibliography

Estellés-Arolas, E. González-Ladrón-de-Guevara, F. (2012). *Towards an integrated crowdsourcing definition*. Journal of Information Science, 14 p.

Shirky, C. (2010). *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Penguin Books. 242 p.

Waibel, G. (2008) You're more social than you think. GNCTPG annual meeting.

Original diaporama can be found online: http://tinyurl.com/ke6cxox