Digital Libraries and Crowdsourcing: A Review

Mathieu Andro, Imad Saleh

Digital Libraries and Crowdsourcing: A Review

Chapter written by Mathieu ANDRO and Imad SALEH, Paris 8, Paragraphe

Abstract

Cataloguing, indexing and correcting the OCR of digitized documents, libraries have often externalized certain activities to service providers with recourse to a low-price workforce in developing countries like Madagascar, India, or Vietnam. From now on, though, they could instead call on the masses of Internet users, that is, crowdsourcing, to realize tasks their own staff cannot handle.

The development of crowdsourcing in libraries is particularly important in the domain of OCR correction. In fact, character recognition software that converts photos of digitized book pages into texts do not provide 100% reliable results and, depending on the quality of the original document, its digitization, its typography and the possible presence of handwritten notes, it may be necessary to correct the texts produced with the help of dictionaries. OCR correction is necessary to enable more efficient whole-text searches of the digitized texts, better referencing of the contents by search engines, the production of eBook in EPUB or MOBI formats so they can be read on

eReaders, data extraction through text mining technologies, or even scientific exploitations related to culturomics. This question of recourse to crowdsourcing is being asked more and more today of libraries, from the very largest of them to the very smallest. In order to bring them part of the solution and bring about an original conceptual contribution to crowdsourcing in libraries, we have written this state of the art, which comes from thesis work.

It will offer conceptual elements to understand this phenomenon, a taxonomy and panorama of the initiatives, and analyses from library and information science points of view.

5.1. The concept of crowdsourcing in libraries

5.1.1. *Definition of crowdsourcing*

Crowdsourcing literally means outsourcing to Internet users, according to Jeff Howe's expression proposed in Wired Magazine in June 2006. According to an authoritative definition, " *Crowdsourcing is a type of participative online activity in which an individual, organization, or company with enough means proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.*" (from Estellés-Arolas, *Towards an integrated crowdsourcing definition*. Journal of Information Science, 2012. [EST 12])

Contrary to these authors, we think crowdsourcing can also exist as participation that is not necessarily and strictly voluntary, as is the case with projects where Internet users contribute by playing games, which are qualified as gamification. We even think crowdsourcing can also call on the involuntary or unconscious participation of Internet users, as is the case, for example, with the reCAPTCHA project. The millions of books digitized by Google Books are OCRized. The words not occurring in dictionaries are then sent to Internet users who, for security reasons, are forced to reassemble jumbled words to prove that they are not robots. In doing this, by creating their accounts on websites, they involuntarily contribute to OCR correction for Google Books and Google Maps. We qualify this involuntary participation of Internet users as implicit crowdsourcing.

Having defined crowdsourcing, all that remains is to explain what it is not. Crowdsourcing must not be confused with outsourcing, for there is indeed a sort of call for bids in the form of a call for contributions; the relationship with the contributor, however, is not contractual. It also must not be confused with "user innovation", as the undertaking remains at the project's initiative, not with the open source since the contribution method is not necessarily collaborative, but can, quite the opposite, appeal to competition.

5.1.2. *Historic origins of crowdsourcing*

This economic model finds its source in

– government appeals to the people to solve scientific problems for recompense starting in the 18th Century;

– competitions and public offerings;

– free service and free access that allowed the consumer to take over part of the producer's work, then the "on-demand" model that allowed him to take over the production decision itself.

Below, we propose a chronology of the historic origins of crowdsourcing and citizen science. This chronology is the result of analyses done on the literature on this subject. It was created using a collection of the most important events at the core of crowdsourcing (call for participation, recompense, collective work, microtasks, outsourcing, wisdom of crowds). The events assembled here are also the most representative of the taxonomy that we propose later in the chapter.

1714 The English government launches a call for scientists to find a solution to determine maritime longitude from a boat. John Harrison, a carpenter and clockmaker, wins the 20,000 pound reward over more than 100 competitors, including Cassini, Huygens, Halley and Newton
1726 A ruling by the King of France, Louis XV, requests that ship captains bring plants and seeds back from foreign countries that they visit
1750 British astronomer Nevil Maskelyne calculates the position of the moon for navigation at sea, thanks to the calculations of two astronomers who made their calculations twice each, and then were verified by a third
1758 Mathematician Alexis Clairaut manages to calculate the orbit of Halley's comet by dividing the calculation tasks among three astronomers
1794 French engineer Gaspard de Prony organizes addition and subtraction microtasks for 24 unemployed barbers in order to develop detailed logarithmic and trigonometric tables
1810 With his new methods of food preservation that will lead to canned food, Nicolas Appert receives 12,000 francs from the French government after a call for contributions
1852 The store "Au bon marché" is the first self-service store. From then on, consumers directly access merchandise without going through the intermediary of a merchant and thus take on part of the producer's work
1857 After a call for volunteer contributions, the Oxford English Dictionary benefits from more than 6 million documents containing word proposals and citations of use
1884 The Statue of Liberty is financed by public donations
1893 During a competition on the livestock market to guess the weight of the cow, Francis Galton notices that the average of a crowd's estimates is closer to the truth than experts' estimates, implying the existence of the "wisdom of crowds"
1895 Librarian James Duff Brown invents free access in libraries. Readers of the Clerkenwell Public Library from then on have direct access to part of the collections
18?? In the field of editing, public offerings multiply to finance the publication of books
1900 The National Audubon Society (USA and Canada) organizes a "Christmas bird count"
1936 Toyota gathers 27,000 people and chooses the best proposed design for its brand logo
1938 In the United States, the Mathematical Tables Project employs 450 out-of-work victims of the economic depression, led by a group of mathematicians and physicists, to tabulate mathematical functions, long before the invention of the computer
195? A Toyota industrial engineer, Taiichi Ōno, invents the "just-in-time" model, predecessor of the "on-demand" model, which would allow production without reserves or unsold articles, just-in-time manufacturing as a function of demand. In a way, it is a matter of outsourcing the production decision itself to the consumer. This model is at the root of on-demand digitization through crowdfunding and on-demand printing
1954 The first telethon in the United States allows fundraising to fight cerebral palsy
1955 The Sydney Opera House is designed and built after a public competition that encouraged ordinary people from 32 countries to contribute to this design project
1979 The Zagat survey (restaurant guide) bases its ratings on a large number of testers. The project was purchased by Google in September 2011
1981 The 3rd edition of the Lonely Planet travel guide is written through the participation of independent travelers
1996 Birth of the Internet Archive
1997 Le livre à la carte: facsimile reproduction of books kept in libraries (on-demand digitization and printing)
1997 Rock band Marillion finances its US tour, thanks to fan donations amounting to $60,000
1998 The Dmoz directory offers content generated by its users. The Web 2.0 is born
2000 Philanthropic crowdfunding platform justgiving.com and the participatory artist financing platform artistshare.com see the light of day. They are followed by multiple initiatives until today
2000 Distributed Proofreader: first participatory book transcription project
2001 Birth of Wikipedia
2003 ESP Game: a game for image indexing
2005 Amazon launches the crowdsourcing platform Amazon Mechanical Turk Marketplace for its own needs and also allows coordination of research societies and institutions and workers on the Web for microtasks
2006 Espresso Book Machine for *in situ* on-demand printing
2006 Jeff Howe proposes the term "crowdsourcing" in Wired Magazine in June 2006
2007 Google Books uses reCAPTACHA to have its untreated OCR corrected by Internet users
2008 The gamification project Fold.it allows advances to be made in the knowledge of proteins, thanks to puzzle games
2011 The Good Judgement Project makes use of Internet users' wisdom of crowds through their geopolitical expectations, which rival those of intelligence experts
2011 Digitalkoot for OCR correction in the form of arcade games
2013 The video game Star Citizen raises a sum of $30,044,586
Table 5.1. *Chronology of crowdsourcing in libraries*

5.1.3. *Conceptual origins of crowdsourcing*

Crowdsourcing finds its conceptual origin in as diametrically opposed ideologies as socialism, libertarianism, humanism or liberalism [AND 14a], where the Californian Ideology would accomplish the most propitious synthesis for the development of crowdsourcing.

Socialist and Marxist ideologies "From each according to his ability, to each according to his needs" production and abolition of the law of value, money is no longer the main motive, free products, spirit of sacrifice, work to serve humanity, socialist emulation, abolition of the fundamental separation of necessary labor and surplus labor, reconsideration of wage labor and social classes, each person able to be employer or employee in turn, overcoming private property through shared use, "collaborative communities", participatory, peer-to-peer contribution economy
Libertarian and anarchist ideologies Critical of the authority of the dominating classes and totalitarianism, direct democracy, equal contribution from the hobbyist and the expert, disappearance of the boundary between producers and consumers, work becomes leisure, weisure (work + leisure) or playbor (play + labor)
Humanism Digital humanities, Internet rehumanization and restitution of humans from a central place on the Web as origin and end, trust in man's abilities superior to those of algorithms, altruism and love of neighbor, concern for the weaker in the face of the strong
Liberalism Outsourcing, integration of the consumer in production, meritocracy, increase in individual freedoms, defense of the freedom of expression, spirit of initiative and enterprise, "fun at work", universalism, internationalism, democracy, invisible hand, spontaneous order (Friedrich Hayek) of Wikipedia, which works through the autonomous action of individuals with no planning, trust in the market, selfemployment, reconsideration of monopolies, "uberization"
Californian Ideology: libertarian liberal philosophy, libertarian philosophy of hippie meritocratic entrepreneurs and philosophy of digerati (digital literati)
Table 5.2. *The conceptual origins of crowdsourcing*

5.1.4. *Critiques of crowdsourcing. Towards the uberization of libraries?*

Crowdsourcing applied to libraries could also be considered a form of library uberization. Uberization could be defined as challenging established societies and professions through the emergence of web platforms allowing non-professionals to offer competing services. In the library domain, it could take the form of replacing the professional, authoritative data producer with a volunteer working for free or even an underpaid worker producing lowquality data outside any legal framework [FOR 11]. This exploitation of the invisible work of the Web's proletariat is sometimes considered "servuction". It is accused of unfair competition by traditional service providers. It would bring about disengagement on the part of workers, like those who employ this interchangeable workforce. It would create impersonal relations and fraud.
As a result, some thinkers, like Bernard Stiegler, talk about creating a "contributory revenue" [STI 15]; others speak of taxing data to return to citizens part of the value that they have created through their invisible data production work; and still others discuss making data produced by the masses joint property.
As for libraries, they sometimes remain too focused on the constitution of collections as a means in and of itself rather than on satisfying the needs of readers. Before mass digitization, they enjoyed a sort of monopoly on access to information, and their administration, forming a corporation with a relative ideological homogeneity, benefitted from prestigious titles of curators. Under these conditions, outsourcing expert work to hobbyists, opening up to the private as a renewed public/private partnership, risks being seen as questioning, losing control, disloyal competition, an attack against social benefits and, finally, an uberization of libraries.
The question of the quality of contributions, the costs connected with monitoring the quality, the individual appropriation of the collective heritage by uneducated laypeople, and the possible malevolence of Internet users will thus be pointed out against crowdsourcing projects that will not be able to
develop without significant change initiatives.

5.2. Taxonomy and panorama of crowdsourcing in libraries

Most of the actors establish a typology of crowdsourcing projects as a function of the public's degree of engagement. In this way, with participatory or contributory crowdsourcing, Internet users are happy to produce data for institutions that come up with projects, pilot their development and frame the public's participation, which remains limited to microtasks only requiring a small individual investment. With collaborative crowdsourcing, Internet users can also interact with one another. Through co-creation, this individual investment is even stronger, as Internet users can actively weigh in on the policy and definition of projects' goals and premises, and sometimes even be the source of the projects themselves.
Beyond this quantitative distinction, we were led to propose a more qualitative taxonomy of crowdsourcing projects in libraries. We distinguish among the following large types:

Explicit crowdsourcing	**Definition**	**Identified projects**
Volunteer crowdsourcing	Recourse to voluntary work from voluntary Internet users	*Participatory uploading and curation*: Oxford's great war archive, Europeana 1914–1918, Internet Archive, Commons Wikimedia, Wir waren so frei, Open Call – Brooklyn Museum, Pin-a-tale, Make history, Click! A Crowd-Curated Exhibition, The Changing Faces of Brooklyn, ExtravaSCANza *Participatory OCR correction*: TROVE, Distributed Proofreader, Wikisource, California Digital Newspaper Collection, Correct, Franscriptor *Participatory manuscript transcription*: Transcribe Bentham, What's on the Menu?, Ancient lives, ArcHIVE, What's the score, Transkribus, les Herbonautes, Do it yourself History, Monasterium Collaborative Archive, Citizen Archivist Dashboard, National Archives Transcription Pilot Project, Field Notes of Laurence M. Klauber, Notes from Nature, Transcribe Bushman, Smithsonian Digital Volunteers Transcription Center *Folksonomy*: Flickr, The Commons, steve.museum, GLAM Wikimedia, Glashelder!, VeleHanden, 1001 Stories Denmark, Historical Maps Pilot, Mtagger, PennTags, Social OAC, Describe me, Tag! You're It!, Freeze tag!, Your Paintings Tagger, Operation war diary
Paid crowdsourcing	Recourse to the work of paid Internet users	*All kinds of work*: Amazon Mechanical Turk Marketplace, 99design, CloudCrowd, Cloud-Flower, CrowdFlower, Upwork, Foule Factory, Freelancer, Guru, Innocentive, ManPower, Mob4hire, MobileWorks, oDesk, Postmates,

		quora.com, Samacource, sparked.com, TaskRabbit, Topcoder, Trada, Turkit, uTest

Implicit crowdsourcing and gamification	**Definition**	**Identified projects**
Implicit crowdsourcing	recourse to involuntary work by Internet users	*OCR correction*: reCAPTCHA
Gamification "human computation" "games with a purpose"	recourse to Internet users' work in game form	*OCR correction*: Digitalkoot, COoperative eNgine for Correction of ExtRacted Text, TypeAttack, Word Soup Game, Smorball, Beabstalk *Indexing*: Art Collector, Google Image Labeler, ESP Game, GWAP, Peekaboom, KissKissBan, PexAce, museumgam.es, Metadata Games, SaveMyHeritage, Picaguess, Wasida
Crowdfunding	recourse to Internet users'financial contributions	*On-demand digitization*: eBook on demand (EOD), books à la carte, Éditions du Phoenix, Chapitre.com, les amis de la BnF, Numalire, revealdigital, Lyrasis, FeniXX, unglue.it, Maine Shared Collections Strategy, International Amateur Scanning League, "Sauvez nos reliures" *On-demand printing*: Espresso Book Machine, Electronic Library, Higher Education Resources ON Demand, Amazon Book Surge, CreateSpace, Jouve, lulu.com, Lightning source, Virtual Bookworm, Wingspan press, iUniverse, Xlibris

Table 5.3. *Taxonomy of crowdsourcing and panorama of the projects for digital libraries*

All of these forms of crowdsourcing shown synthetically and introductorily in the above tables are the object of an analysis developed in the section that follows and returns to this taxonomy's structuration.

5.2.1. *Explicit crowdsourcing*

5.2.1.1. *Volunteer crowdsourcing*

This is the most obvious and classic form of crowdsourcing, but recourse to volunteers could quickly reach its limits faced with the proliferation of projects. Furthermore, nothing indicates that future generations of pensioners, who are sometimes a significant portion of contributors, will have the same interests.

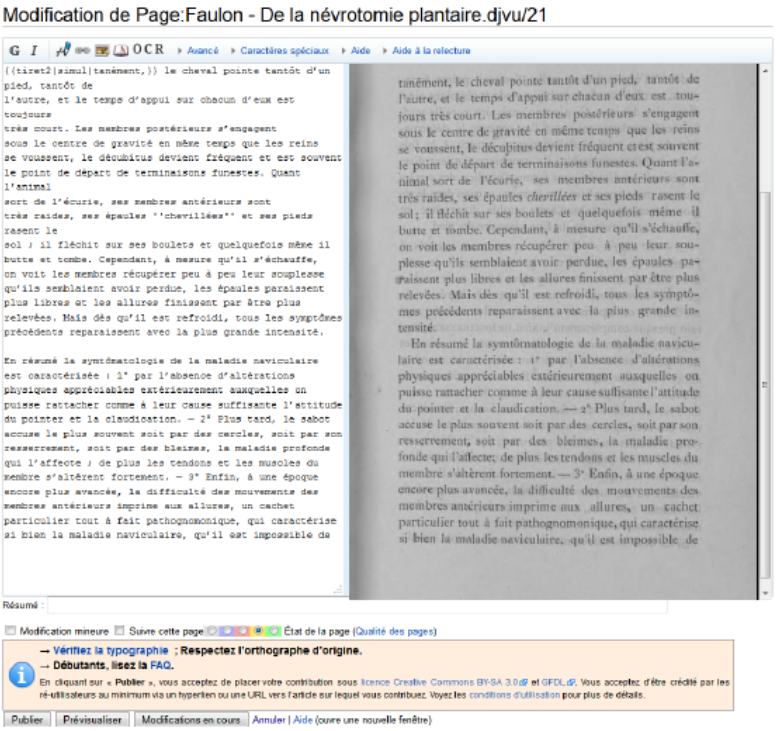

Figure 5.1. *Page from an old thesis saved at the National Veterinary School of Toulouse for which OCR correction is proposed via Wikisource*

5.2.1.2. *Paid crowdsourcing*

The primary users of the largest paid crowdsourcing platform, Amazon Mechanical Turk Marketplace, are American research laboratories. This platform brings together those offering and those seeking online work, generally in the form of microtasks. With it, crowds of workers worthy of the largest multinationals, with diverse profiles, among more than 500,000 Internet users permanently available in nearly 200 countries, particularly the USA and India, are to be recruited in a few minutes time, without administrative procedures, at costs freely determined by supply and demand [IPE 10]. It thus allows the realization of jobs that would have required years of thankless before, done by "burn outs", in half a day. As for the workers on the platform, they are free to work where they want, when they want, as much as they want, for whom they want, based on their own interests, to be employer and employee in turn, and to work for a client rather than a boss.

The name "Amazon Mechanical Turk Merketplace" is cleverly inspired by an automatic chess player thought up in the 18th Century that was supposedly gifted with artificial intelligence when, in fact, a human was hidden behind it. In the same way, behind the results that are believed to be done by pwerful algoritms, there may, in fact, the crowds of hidden humans, particularly through paid crowdsourcing.

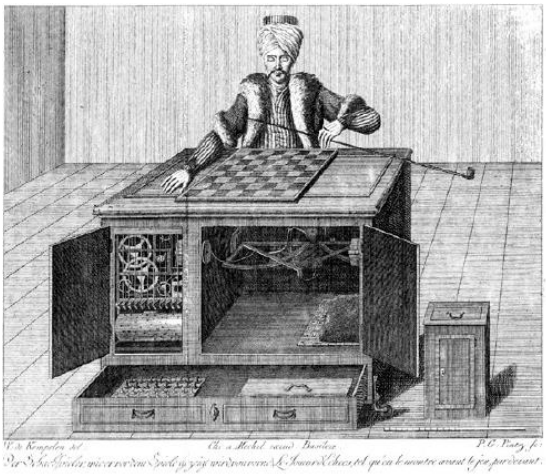

Figure 5.2. "Türkischer Schachspieler" by Karl Gottlieb von Windisch. 1783. Public domain via Wikimedia Commons

5.2.2. *Gamification and implicit crowdsourcing*

The contributor's will is not necessarily the primary goal for participants in these forms of crowdsourcing. They call on Internet users' desire to play to receive work from them (gamification) or make them work without their knowing it (implicit crowdsourcing).

5.2.2.1. *Gamification*

Gamification consists in making Internet users work through games with a useful and productive end ("games with a purpose"). It could be defined as the act of applying design, psychology and video game elements in other contexts [DET 11].

The simple act of giving points for Internet users' participation therefore must not be confused with gamification, but rather results in a sort of "pointification". Gamification is also different from "serious games" because it does not aim to educate for personal development, but rather to achieve goals outside oneself like correcting OCR or indexing digitized photographs [AND 15b].

Unlike explicit crowdsourcing, doing randomly performed, out-ofcontext microtasks in a game is generally less favorable to personal development and the acquisition of knowledge, but it could allow work that is sometimes rather thankless to be done more easily.

Figure 5.3. *Screenshot of the Digitalkoot OCR correction game [CHR 11]* Digital Libraries and Crowdsourcing: A Review 13

5.2.2.2. *Implicit crowdsourcing*

The notion of implicit crowdsourcing was conceptualized by [HAR 13], but the term is still not very widespread in the literature. The Internet users who participate do it involuntarily or unconsciously. Implicit crowdsourcing could thus be considered less ethical than explicit, voluntary crowdsourcing by unpaid workers on the Web, an intrusion of eCommerce seeking to instrumentalize Internet users.

The most emblematic project of this kind of crowdsourcing in the domain of digital libraries is reCAPTCHA. In order to create an account on a website and avoid any attack by robots, the websites require Internet users to reassemble jumbled words, thereby proving that they are not malicious bots. The programs Google Books and Google Maps have thus cleverly used this system to have untreated OCR from their campaigns of digitization by masses of Internet users corrected by comparing their input. Thus, 200 million words would be compared each day, 12,000 h of volunteer work collected and, according to our calculations, 146 million euros per year saved by Google through text correction services [AND 15a].

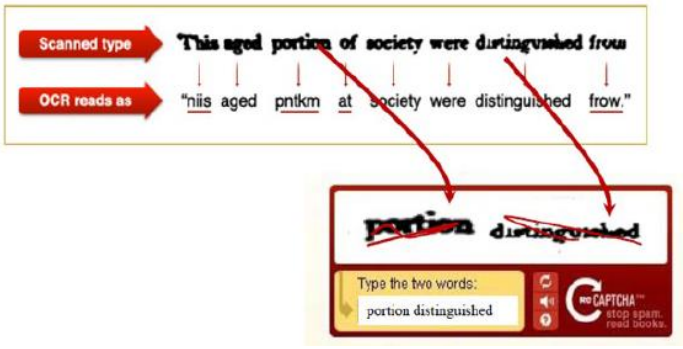

Figure 5.4. *Diagram explaining how reCAPTCHA works according to https://www.google.com/recaptcha*

5.2.3. *Crowdfunding*

Sometimes considered institutional begging [AYR 13], crowdfunding is indeed a form of crowdsourcing that calls not on the work of Internet users, but on their financial resources. In libraries, it can be used to acquire documents or to finance digitization.

5.2.3.1. *On-demand digitization*

On-demand digitization allows libraries to offer digital reproduction services by having Internet users support the costs, outsourcing part of the costly, thankless task of selecting documents that still deserve to be digitized and obviously completing their digitization programs. The user thus finds himself placed at the center of library policy [GST 11], whereas in the past, libraries sometimes tended to neglect them by principally focusing on their collections. The documentary policy of the digital library thereby becomes a co-construction between the librarians and the general public since the acquisition policy is from then on shared. For Internet users, on-demand digitization gives them access to a digital reproduction service. For potential patrons and investors, on-demand digitization could be the chance to finance the digitization of books that may interest such and such audience, and eventually to collect a return on investments through web traffic created by these books on the advertisement model Google Adwords.

This economic model could allow public funds, which have become more rare, to be concentrated on the digitization of documents with patrimonial, historic and scientific interest but not interesting private sector and to allow private funds from individuals or patrons finance the digitization of works interesting the general public or communities of scholars. In doing this, libraries would have a chance to better refocus on their own areas of expertise and better value the skills of curators in the patrimonial, historic and scientific domain.

On-demand digitization by crowdfunding is a new form of public subscription allowing new life to be given to a work. It is particularly well adapted to the current situation, where only leftovers still need to be digitized after large mass-digitization programs pass through.

The main difficulty of on-demand digitization projects consists in automatically evaluating the costs of document digitization. In fact, it is claimed that for these projects, a cost estimate is necessary for each demand. This estimate serves to evaluate the cost of digitization. Producing this estimate requires verifying the presence of the document, its state, its actual page count, its format and how wide it opens, all of which will determine how many pages must be digitized and the type of scanner to be used, thus the cost of digitization. Unfortunately, after the Internet user receives the estimate, he only very rarely proceeds with his demand through an order, as his desire to purchase has been surpassed and he may be surprised by the cost to be paid. At the end of the day, the work time spent producing estimates costs as much as the money collected from Internet users, and in the absence of an automatic calculation for digitization costs or a subsidy through public funds, on-demand digitization projects are hardly ever viable [AND 14b].

5.2.3.2. *On-demand printing and libraries*

Although it is not a matter of crowdsourcing, the economic model is identical to that of on-demand digitization, from which is it often indistinguishable. Here we are dealing with the revival of a print through the digitized document. This model, more and more often used in publishing to produce just in time, without reserves or unsold articles, has been applied to library digitization programs [AND 15c].

As we have seen with on-demand digitization, the documentary policy of digital libraries and the constitution of digital collections are henceforth more of a co-construction; furthermore, they are henceforth partly the work of Internet users. With on-demand printing, we could even go so far as to imagine a physical library directly made up of prints demanded by its readers, printed in a few minutes through an Espresso Book Machine and, after being returned by the reader, constituting a collection built by the user and made up of works having all been consulted at least once [LEW 10]. This way of functioning would be radically different from the acquisition of libraries, which currently depends essentially on the anticipation of needs

and the purchase of books in case they one day interest a reader, a policy that is thus not exclusively focused on the user. In this way, it extends the possibility already offered by libraries to their readers to suggest acquisitions.

Figure 5.5. *The Espresso Book Machine according to http://ondemandbooks.com*

Beyond this taxonomy, there are forms of crowdsourcing that have probably not yet been invented, such as gamification paid according to the results obtained through playing, the resource to citizens, possibly paid, for themselves digitizing documents within libraries, or even the application of a reCAPTCHA benefitting public libraries or a reCAPTCHA charging for the OCR corrected and sharing its profits with the sites that accepted to implement it.

5.3. Analyses of crowdsourcing in libraries from an information and communication perspective

5.3.1. *Why do libraries have recourse to crowdsourcing and what are the necessary conditions?*

Clay Shirky thought that if Americans spent their time on projects like Wikipedia instead of watching television, they could create 2,000 projects on the same scale as the famous participatory encyclopedia [SHI 10]. As for Luis Von Ahn, he claimed that the 425 million images on Google Images could have been indexed in just 31 days by 5,000 Internet users playing the ESP Game [VON 08]. Whatever it is, there would be a significant reserve of good will that libraries could benefit from, especially as they already have experience in motivating communities, the setting of these good wills; they have a good image with the public to whom they seem worthy of trust and they seem to serve general interests and whom they could, consequently, more easily recruit.

They could therefore have recourse to crowdsourcing, that is, to outsourcing microtasks to masses of Internet users to reduce their costs or to multiply their human resources and realize a painstaking, tedious task that they do not have the means to take on, or even to complete, undertake, or make possible projects that until now were unachievable, impossible or even unimaginable [HOL 10]. By taking advantage of the collaboration of Internet users, libraries could benefit from limitless knowledge and skills, far beyond that of their limited teams, all despite the excellent general education of their directors. Crowdsourcing therefore challenges the borders of the organization, as it allows value to be created beyond its borders [REN 14]. Libraries could thus tap into the greater strength lying beyond their organizations and recruit such and such scholar or specialist to identify a location or person in a photograph, recognize a book cover, date an object in a curio cabinet, etc. Libraries could thus get engaged in a participatory redocumentarization process and see their collections revisited and reinvented. Beyond these incentives, libraries could also seek to deeply engage the user in their collections, to democratize heritage conservation in the form of an "empowerment", an emancipation, and a change in relations between heritage and society in the name of a right to information and participation, to improve their image, to seek to innovate or, unfortunately, also to start institutional communication around a trendy subject.

Although it is absolutely possible to talk about an authentic use of data produced by users, there seem to be two different conceptions in libraries. One thinks that libraries need Internet users, while the other thinks that Internet users need libraries. The first is utilitarian and economical. It humbly recognizes its need for reinforcements and is truly seeking to produce a result, thanks to collaboration between heritage and society. The second is more strongly tied to democratic considerations and seeks instead to practice crowdsourcing for the sake of practicing crowdsourcing, publicizing the process and increasing public participation as an end in itself. While this second conception holds sway, the work produced by Internet users is barely valued or used and the metadata that they seize will only rarely be integrated into library information systems, which may discourage and be seen as a betrayal by volunteers, as we will see in a later chapter.

Whatever the case may be, libraries will only be able to get engaged in the crowdsourcing path under the condition that the tasks concerned can be performed online as microtasks, that they do not involve confidential data, that they can be undertaken independently and requiring little interaction, education and communication, and, finally, that they can be accomplished by non-specialized amateurs.
There is a growing interest for crowdsourcing in libraries in the world literature, as illustrated in Figure 5.6.

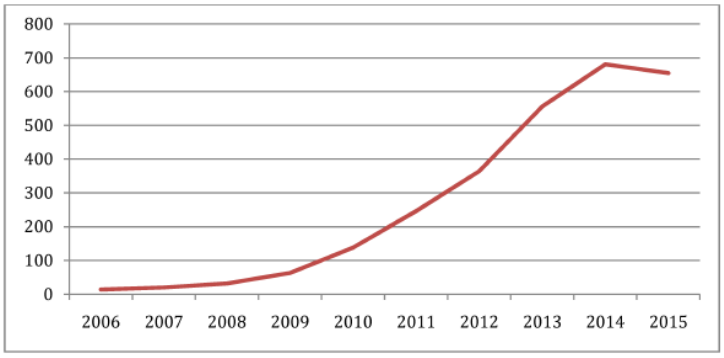

Figure 5.6. *Number of publications indexed in Google Scholar as a function of their years of publication and responding to the search "crowdsourcing AND library AND digitization"*

5.3.2. *Why do Internet users contribute? Taxonomy of Internet users' motivations*

Among the many motivations that lead Internet users to contribute to crowdsourcing projects in libraries are mainly seen intrinsic motivations pushing the individual to act selflessly and for the pure joy that the work brings him and the extrinsic motivations pushing him to work for the effects and results obtained, thanks to this work, like recognition, recompense or remuneration. We have identified and organized the motivations shown in Figure 5.8 from the literature.

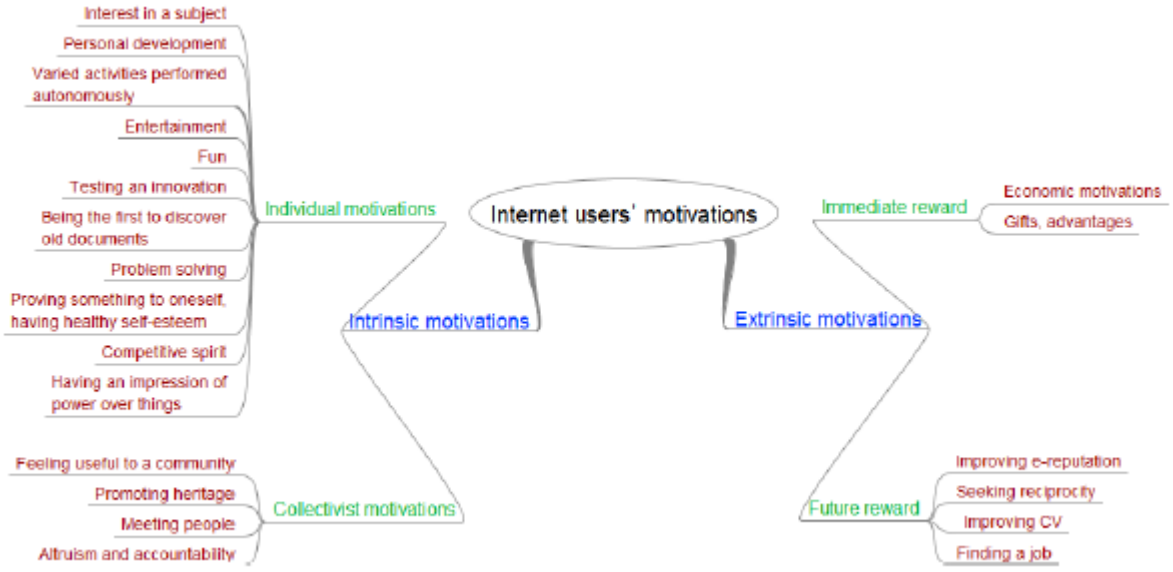

Figure 5.8. *Taxonomy of the motivations of Internet users who participate in crowdsourcing projects in libraries*

The kinds of motivations can be very diverse as a function of the types of crowdsourcing types of individuals and cultures, so if volunteer crowdsourcing encourages Internet users to seek personal development, gamification is rather appealing to the need for useful entertainment. The emergence of extrinsic motivations and particularly symbolic recompense, real or virtual, can sometimes take place to the detriment of the quality of data produced and of intrinsic values, that is, those that lead them to act out of pure interest in the job itself. Starting in 1975, an experiment by Edward L. Deci showed that if

people were remunerated for doing puzzles, they lost all interest in these activities if they no longer saw rewards for doing it.

5.3.3. *From symbolic recompense to concrete remuneration*

As the preceding concept map of motivations shows, crowdsourcing projects are necessarily carried out for the mutual benefit of the institution and the Internet user. In addition to the intrinsic movements, the rewards can range from symbolic rewards (ranking, grades, medals) to very real rewards, from gifts even to remuneration. As such, volunteers with the Foldit project were publicly thanked in an article in the celebrated academic journal *Natural Structural & Molecular Biology* vol. 18, 2011 for having made the discovery of a very important enzyme's structure possible. Other volunteers with other projects were mentioned in newsletters and invited to talk about their work at conferences, and they were rewarded training courses, subscriptions, books, T-shirts, MP3 players, gift certificates, tours or trips.

5.3.4. *Communication for recruiting contributors*

Cultural institutions benefit from a good public image and seem worthy of trust and to serve public interests. As a result, they have solid advantages for recruiting volunteers. Among the communication means used for crowdsourcing projects, we could mention campaigns in the associative, local, national and trade press, the publication of articles and posters, the distribution of leaflets, putting up stickers and posters particularly for conferences and symposia, the organization of public meetings or specific events, and radio and television presence, but also the production of videos, the use of social networks, forums, mailing lists, direct mail campaigns, institutional websites, and, finally, the purchase of specific words in the Google Adwords campaign.

A crowdsourcing site must always have a homepage that describes the project simply and clearly explains its end goal and progress, and immediately invites volunteer participation by showing them how their participation will be useful and how they will be guided and recognized [MCK 15].

5.3.5. *Community management for keeping contributors*

The majority of the data produced in the framework of crowdsourcing projects has been produced by a well-determined minority of participants and not by anonymous masses.

As seen in the previous diagram where each square corresponds to the contributions of a single person and where the size of each square is proportionate to quantity of contributions, all volunteer crowdsourcing projects also show us that the largest part of contributions is the result of a minority and that it is thus not really a matter of anonymous crowds, but rather of a well-defined community of volunteers [OWE 13]. Under these conditions, it would therefore be more judicious to speak of communitysourcing [CAU 12] or even nichesourcing and to seek to recruit well-targeted people rather than addressing faceless crowds.

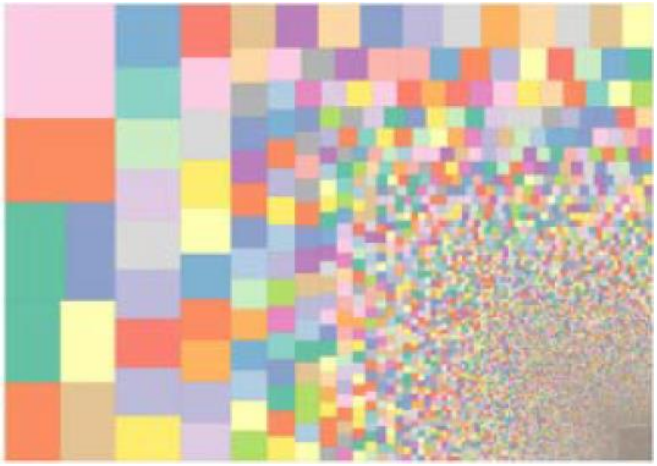

Figure 5.9. A few Internet users produce the largest part of contributions (according to Brumfiels's blog manuscripttranscription.blogspot.fr in 2013)

If the standard profile of a volunteer is a college-educated man with a high level of studies, in an elevated socio-professional category and having just finished his studies or who is retired, the communities of contributors also include more diverse profiles and form communities of pairs having similarities and common goals whose dynamism must be maintained by community management. In fact, the volunteers will have to be recruited by communication acts, supervised, managed, supported by hotlines, helpdesks, forums, educated with manuals, tutorials, motivated, regularly updated, their contribution moderated, their quality controlled, the participation statistics followed, and finally, the produced data reintegrated.

5.3.6. The quality and reintegration of produced data

The issue of quality and vandalism is a central argument of opponents of outsourcing data production to amateur Internet users. There are, however, many proven ways to guarantee one and prevent the other using robots, self- regulation, inspection by professionals or volunteers, evaluations, votes, aptitude tests, or even double-keying.
Whatever the case may be, there are numerous studies that show that data produced by those who consult a digitized document online can be of the highest quality, as the person consulting the document is generally someone who knows the subject covered in the consulted document well. Furthermore, as expected by followers of the "wisdom of crowds", the diversity of profiles constituted by a crowd of Internet users and the "law of large numbers" would have the effect of neutralizing individual errors in the mass of accurate data [BOE 12] and sometimes even providing better results than those obtained by experts. This seems to be confirmed by several comparative studies [ROR 10, OOM 11].
However, the data produced by Internet users are not always reintegrated by institutions sometimes aiming more towards democratization of heritage or even institutional communication on a trendy subject than real Internet user participation. When they are reused, these data born of free, volunteer contributions should always be distributed under legal conditions authorizing the largest possible reuse.

5.3.7. *The evaluation of crowdsourcing projects*

Although it has proven difficult to collect figures in the literature except for [CAU 12], it seems that crowdsourcing projects are far from all being profitable in the sense that the profits in data harvested are not always on the level of the costs for project management, platform development, administration, maintenance, hosting, communication, community management and reintegration of produced data.
In a previous study [AND 15a], we had estimated that the California Digital Newspaper Collection project gathered more than 1,500 € of OCR correction work per month, the Digitalkoot project more than 2,000 € per month, the TROVE project nearly 35,000 € per month and the reCAPTCHA more than 12 million euros per month.
The projects that seem to work the best are those that have clear ends goals, lead an efficient communication campaign, and have managed and motivated communities at their disposal. Those that fail generally call on overly complex tasks, overly specialized knowledge requiring too great an investment in training, do not communicate sufficiently with the volunteers [RID 13], and sometimes neglected internal change management.

5.4. Conclusions on collective intelligence and the wisdom of crowds

If we average the individual estimates of a crowd concerning the weight of a cow, the number of marbles in a jar, the temperature in a room or the response to a general culture question on a game show like "Who Wants To Be A Millionaire?", it will become clear that this average is closer to the truth. This phenomenon was identified long ago as "vox populi vox dei" and well understood by Machiavelli, who wrote:
"*I say that the people are more prudent and stable, and have better judgment than a prince; and it is not without good reason that it is said, 'The voice of the people is the voice of God'; for we see popular opinion prognosticate events in such a wonderful manner that it would almost seem as if the people had some occult virtue, which enables them to foresee the good and the evil. As to the people's capacity of judging things, it is exceedingly rare that, when they hear two orators of equal talents advocate different measures, they do not decide in favor of the best of the two, which proves their ability to discern the truth of what they hear*." [MAC 37]
Today, this phenomenon is known as the "wisdom of crowds" [SUR 04]. Many crowdsourcing projects rely on this phenomenon to obtain quality data or sometimes even true expertise. As such, an American intelligence agency, the "Good Judgment Project", relies on the geopolitical forecasts of crowds of Internet users quantitatively estimating the probability or improbability of such and such event. In the same way, big data and the analysis of geographic locations occurring alongside the name "Ben Laden" in the international press using text-mining technologies has shown that those locations were near where he was hiding. When data form crowds, they could therefore also form science, and thanks to crowdsourcing, human brains could be connected like high-powered processors contributing to a calculation much more massive than that of algorithms. We could thus speak of "Human computation" [VON 06].
However, in history, crowds are not always distinguished by their wisdom, but sometimes rather by the criminal irresponsibility of the individuals making up the masses, as Gustave Le Bon brought up in his *The Crowd: A Study of the Popular Mind*, particularly in light of the events tied to terror in the French Revolution, and this long before the totalitarian experiences of the 20th Century. On the Web today, while the madness of the masses does not have the same breadth, the rumors, conspiracy theories and collective paranoia sometimes also seem to disprove the existence of any wisdom of the crowds.
But beyond the simple production of data, recourse to crowds of amateurs with diverse profiles not seeking to reproduce the established models with which professionals have been educated can also be a source of happy coincidences (serendipity), "unexpected readers", accidental discoveries or even innovative breakthroughs. In any case, it encourages the development of an ecosystem of innovation. Thus, according to Eric Von Hippel, "user innovation" theorist, 46% of American companies in innovating sectors find their origins in a simple do-it-yourself consumer [VON 05].

5.5. Bibliography

[AND 14a] ANDRO M., SALEH I., "Bibliothèques numériques et crowdsourcing: une synthèse de la littérature académique et professionnelle internationale sur le sujet", in ZREIK D.K., AZEMARD G., CHAUDIRON S. *et al.* (eds), *Livre postnumérique: historique, mutations et perspectives*. *Actes du 17e colloque international sur le document électronique (CiDE.17)*, Fès, Morocco, 2014.

[AND 14b] ANDRO M., RIVIÈRE P., DUPUY-OLIVIER A. *et al*.,"Numalire, une expérimentation de numérisation à la demande du patrimoine conservé par les bibliothèques sous la forme de financements participatifs (crowdfunding)", *Bulletin des Bibliothèques de France*, contribution du 2 octobre, 9 p., 2014.

[AND 15a] ANDRO M., SALEH I., "La correction participative de l'OCR par crowdsourcing au profit des bibliothèques numériques". *Bulletin des Bibliothèques de France*, contribution du 16 juin, 8 p., 2015.

[AND 15b] ANDRO M., SALEH I., "Bibliothèques numériques et gamification: panorama et état de l'art", *I2D – Information, données & documents*, vol. 52, no.4, pp. 70–79, 2015.

[AND 15c] ANDRO M., KLOPP S., "L'impression à la demande et les bibliothèques", *Bulletin des Bibliothèques de France*, contribution du 13 février, 7 p., 2015.

[AYR 13] AYRES, M.-L., Singing for their supper': Trove, Australian newspapers, and the crowd, *IFLA World Library and Information Congress*, Singapore, 2013.

[BOE 12] BOEUF G., ALLAIN Y.-M., BOUVIER M., "L'apport des sciences participatives dans la connaissance de la bioviversité", rapport remis à la Ministre de l'Ecologie, 2012.

[CAU 12] CAUSER T., WALLACE V., "Building a volunteer community: results and findings from transcribe bentham", *Digital Humanities Quarterly*, vol. 6, no. 2, 26 p., 2012.

[CHR 11] CHRONS O., SUNDELL S., "Digitalkoot: making old archives accessible using crowdsourcing", *HCOMP 2011: 3rd Human Computation Workshop*, San Francisco, CA, 2011.

[DET 11] DETERDING S., DIXON D., KHALED R. *et al.*, "Gamification : toward a definition", *ACM CHI Gamification Workshop*, New York, NY, 2011.

[EST 12] ESTELLÉS-AROLAS E., GONZALEZ-LADRON-DE-GUEVARA F., "Towards an integrated crowdsourcing definition", *Journal of Information Science*, vol. 38, no. 2, 14 p, 2012.

[FOR 11] FORT K., ADDA G., COHEN K.B., "Amazon mechanical Turk: gold mine or coal mine? ", *Computational Linguistics*, vol. 37, no. 2, pp. 413–420, 2011.

[GST 11] GSTREIN S., MÜHLBERGER G., "Producing eBooks on demand: a European library network", 2011.

[HAR 13] HARRIS C.G., Applying human computation methods to information science, PhD dissertation, University of Iowa, 2013.

[HOL 10] HOLLEY R., "Crowdsourcing: how and why should libraries do it?", *D-Lib Magazine*, vol. 16, nos 3–4, 2010.

[IPE 10] IPEIROTIS P.G., "Demographics of mechanical Turk", 2010.

[LEW 10] LEWIS D.W., "The user-driven purchase giveaway library", *Educause Review*, vol. 45, no. 5, pp. 10–11, 2010.

[MAC 37] MACHIAVEL N., *OEuvres complètes*, Tome premier, Auguste Desrez, Paris, 1837.

[MCK 15] MCKINLEY D., "Heuristics to support the design and evaluation of websites for crowdsourcing the processing of cultural heritage assets", 2015.

[MOI 13] MOIREZ P., MOREUX J.P., JOSSE I., "Etat de l'art en matière de crowdsourcing dans les bibliothèques numériques", Livrable L-4.3.1 du projet de R&D du FUI 12 pour la conception d'une plateforme collaborative de correction et d'enrichissement des documents numérisés, 2013.

[OOM 11] OOMEN J., AROYO L., "Crowdsourcing in the cultural heritage domain: opportunities and challenges", 5th International Conference on Communities & Technologies, Brisbane, Australia, June–July 2011.

[OWE 13] OWENS T., "Digital cultural heritage and the crowd", *Curator: The Museum Journal*, vol. 56, pp. 121–130, 2013.

[REN 14] RENAULT S., "Crowdsourcing : La nébuleuse des frontières de l'organisation et du travail", *RIMHE: Revue Interdisciplinaire Management, Homme(s) & Entreprise*, no. 11, pp. 23–40, 2014.

[RID 13] RIDGE M., "From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing", *Curator: The Museum Journal*, vol. 56, no. 4, pp. 435–450, 2013.

[ROR 10] RORISSA A., "A comparative study of Flickr tags and index terms in a general image collection", *Journal of the American Society for Information Science and Technology*, vol. 61, no. 11, pp. 2230–2242, 2010.

[SHI 08] SHIRKY C., *Here Comes Everybody: The Power of Organizing without Organizations*, Penguin Books, London, 2008.

[SMI 11a] SMITH-YOSHIMURA K., SHEIN C., *Social Metadata for Libraries, Archives and Museums Part 1: Site Reviews*, OCLC Research, 2011.

[SMI 11b] SMITH-YOSHIMURA K., GODBY C.J., HOFFLER H. *et al.*, "Social metadata for libraries, archives, and museums: survey analysis", OCLC Research, 2011.

[SMI 12a] SMITH-YOSHIMURA K., "Social metadata for libraries, archives, and museums: executive summary", OCLC Research, 2012.

[SMI 12b] SMITH-YOSHIMURA K. HOLLEY R., "Social metadata for libraries, archives, and museums: recommendations and readings", OCLC Research, 2012.

[STI 15] STIEGLER B., *La société automatique. 1, l'avenir du travail*, Fayard, Paris, 2015.

[SUR 04] SUROWIECKI J., *La sagesse des foules*, traduction de the wisdom of crowds, J.-C. Lattès, Paris, 2004.

[VON 06] VON AHN L., "Games with a purpose", *IEEE Computer Magazine*, vol. 39, no. 6, pp. 96–98, 2006.

[VON 08] VON AHN L., DABBISH L., "Designing games with a purpose", *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[VON 05] VON HIPPEL E., *Democratizing Innovation*, MIT Press, Cambridge, 2005.