Digital Libraries and Crowdsourcing

Digital Tools and Uses Set

coordinated by
Imad Saleh

Volume 5

Digital Libraries and Crowdsourcing

Mathieu Andro

iSTE

WILEY

First published 2018 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2017958934

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-161-1

Contents

Preface

In lieu of outsourcing certain tasks to service providers with access to countries where labor is cheap, libraries throughout the world are relying more and more on groups of internet users, turning their relationship with users into one that is more collaborative. After a conceptual chapter about the consequences of this new economic model on society and on libraries, an overview of projects in the areas of on-demand digitization, participative correction of OCR especially in the form of games (gamification) and folksonomy will be presented. This panorama leads to an overview of crowdsourcing applied to digitization and digital libraries and analyses in the area of information and communication sciences.

Acknowledgments

I would like to thank Imad Saleh, Professor at the Paragraphe Laboratory of Paris 8 University, for having agreed to supervise my thesis project, for his kindness and for his advice throughout the entire project; Samuel Szoniecky, Senior Lecturer at the Paragraphe Laboratory of Paris 8 University, for having agreed to be the co-director of my thesis and for having invited me to speak with his students; Ghislaine Chartron (Professor at the National Conservatory of Arts and Crafts); Stéphane Chaudiron (Professor at Charles-de-Gaulle University, Lille 3); Céline Paganelli, Senior Lecturer – HDR (accreditation to supervise research) at Paul Valéry University, Montpellier; Alain Garnier, CEO of Jamespot and crowdsourcing advisor at the Groupement Français des Industries de l'Information (GFII) for having agreed to be an examiner of my thesis; François Houllier, Institut University, National de la Recherche Agronomique (INRA), for letting me participate alongside him in a task force on citizen science in order to submit a report on

the subject at the request of the appropriate ministers; Odile Hologne, from the department of promoting scientific and technical information of the Institut National de la Recherche Agronomique (INRA), for having encouraged experimentation around INRA's Numalire project within the framework of my work; Filippo Gropallo and Denis Maingreaud, from the company Orange and the company Yabé, for their project Numalire in which they allowed me to participate, and for their collaboration throughout this research project; Marc Maisonneuve and Emmanuelle Asselin, from the consulting firm TOSCA, for their collaboration on the book that we published together on software and platforms for developing digital libraries; Gaëtan Tröger, Ecole des Ponts ParisTech, for his collaboration in the study that we carried out on the visibility and statistics of the consultation of digital libraries; Pauline Rivière, Sainte-Geneviève Library, and Anaïs Dupuy-Olivier, Académie de Médecine, for their collaboration in the feedback on the Numalire experiment that we wrote together; Robert Miller, Internet Archive, for the collaboration that we had at Sainte-Geneviève Library, which became the first library in France to participate in the Internet Archive; Stéphane Ipert, Centre de Conservation du Livre, for the collaborations and interesting discussions that we had; Pierre Beaudoin and Rémi Mathis, previous and current presidents of Wikimedia France, an association with which collaborations with Wikisource were achieved (National Veterinary School of Toulouse in 2008) or only envisioned (Sainte-Geneviève Library); Valérie Chansigaud, science historian and Wikipedia contributor, with whom first contact was established at the museum followed by a pilot experiment in digitization and participative correction of OCR, which was conducted in 2008 at the National Veterinary School of Toulouse; Gilonne d'Origny, from the company ondemandbooks.com, with whom a collaboration on the first installation of an Espresso Book Machine in France was unsuccessful; Daniel Teeter, from the company Amazon, for the interesting opportunity for partnership that was nearly established; Juan Pirlot de Corbion, founder of chapitre.com and YouScribe, for the passionate discussions that we had over the course of our meetings; Daniel Benoilid, founder of the paid crowdsourcing company Foule Factory, for the discussions that we had; Jean-Pierre Gerault, CEO of the company I2S, leader in the area of manufacturing scanners for the digitization of heritage, president of the Comité Richelieu and CEO of Publishroom, for the interesting discussions that we had; Arnaud Beaufort, National Library of France, whom I met during the Wikimedia days at the National Assembly and with whom I then had an interesting conversation; Silvia Gstrein and Veronika Gründhammer, University of Innsbruck, for having invited me to speak at the Ebooks on Demand 2014 conference; Yves Desrichard and Armelle de Boisse, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques, for having allowed me to speak during the "Quoi de neuf en bibliothèques ?" days these last 5 years; Thierry Claerr, Ministry of Culture and Communication, who allowed me to speak regularly at the ENSSIB and sought me out to write a collaborative work, and with whom I had some very enriching discussions; Jean-Marie Feurtet, Agence Bibliographique de l'Enseignement Supérieur, for our collaboration on a mutualization project of a digital library and for having invited me to speak at the 2011 ABES; Nicolas Turenne, Institut National de la Recherche Agronomique (INRA), for having invited me to show the preliminary results of this work at the seminar entitled "Digital Traces" (Cortext group, Institute for Research and Innovation in Society); Pierre-Benoît Joly, director of the Institute for Research and Innovation in Society (IFRIS), for having invited me to give a master's level course in Digital Studies and Innovation (NUMI); SNCF for the comfort of the train trips I took while writing this thesis; Google for the Google Drive service, which was used to write the thesis while providing real-time access to it for the director, my collaborators and my contacts who then had the opportunity to add comments; my wife Véronique and my three children Terence, Orégane and Eloïse.

I also want to thank the following people for the constructive comments that they added to the text of the thesis made available in its first draft on Google Drive: Christine Young (proofreading the article in English), Wilfrid Niobet (one idea, eight comments, six corrections), Célya Gruson-Daniel (three comments, four corrections), Olivia Dejean (nine corrections), Michaël Jeulin (seven corrections), Catherine Thiolon (ten comments), Caroline Dandurand (five comments), Diane Le Hénaff (three comments), Sophie Aubin (two comments), Nicolas Ricci (one comment), Pauline Rivière (one comment), Frédérique Bordignon (one comment), Sylvie Cocaud (one comment), Marjolaine Hamelin (one comment), Silvère Hanguehard (one comment), Christine Sireyjol (one comment), Odile Viseux (one comment), Véronique Decognet (one comment), Dominique Fournier (two corrections) and all of the "unknown soldiers" who remained anonymous in their comments (82 corrections).

Mathieu ANDRO
November 2017

Introduction

Libraries already resort to outsourcing certain tasks involved in entering bibliographic records, cataloguing, indexation or OCR correction, to service providers in countries where labor is inexpensive. This outsourcing has remained within a contractual and limited framework and has not profoundly overturned the underlying ways in which libraries work. However, with the development of crowdsourcing, it is possible to imagine externalizing (outsourcing) some of these tasks not to service providers but to "crowds" of Internet users and therefore having amateurs carry out some of the professionals' work. Crowdsourcing thus changes the paradigm up on which libraries are based, which now largely centers around the creation and conservation of collections. It also changed the relationship between the service providers, namely the librarians, and their consumers, namely the users. The latter are also becoming active producers of services. Crowdsourcing could also interrogate the collection management policies of libraries, which anticipate need based on a supply that is not directly or immediately determined by demand. This is especially the case with the on-demand digitization by crowdfunding, a form of crowdsourcing that calls not on the work of crowds, but on their financial resources, or with the printing on demand which is inseparable from it. With these on-demand economic models, the collection management policy is finally shared with users who decide what will be digitized and/or printed. In this way, the collections become the work of the users.

This book has the goal of providing responses to the question of relying on crowdsourcing for library professionals, as well as for students, researchers in information and communication sciences and, more generally, people interested in collective intelligence projects. It is the result of a thesis on information and communication sciences that simultaneously includes action research, an experiment and an analysis of the literature [AND 16]. This thesis itself has previously been the subject of an article using the main contributions [AND 17].

Beyond the questions of costs/benefits and advantages/disadvantages, the question of an evolution of the librarian's profession refocused on their singular skills will be addressed. This work also has the scientific goal of providing a contribution to knowledge of crowdsourcing on the theoretical and conceptual level around economic models.

This work is limited to the application of crowdsourcing in the area of digitization and digital libraries. Since the 1990s, the digitization of documents has been widespread in libraries. Today, with mass digitization and the development of gigantic digital libraries such as Google Books, which has crossed the threshold of 30 million books, or Internet Archive, Hathi Trust, Europeana, the "harvester" of European digital libraries, it is becoming more and more difficult to identify printed matter that has not been digitized and still deserves to be, among the 130 million[1] existing titles printed since the invention of printing.

A significant part of what has been digitized by libraries has never been put online. It generates duplicate digitization and is "sleeping" on CD-ROMs, DVDs or external hard drives whose lifetime is limited. The development of a digital library can, in fact, be expensive in terms of software administration and servers, and the result can be disappointing in terms of functionalities, durability, costs and visibility. In 2012, we published a study dedicated to the software programs YooLib (Polinum), Invenio (CERN), ORI-OAI (universities), DSpace (DuraSpace), DigiTool (Ex Libris), Mnesys (Naoned), ContentDM (OCLC), Eprint (University of Southampton), Greenstone (University of Waikato) and Omeka (George Mason University) [AND 12]. In this study, we found that it was more advantageous for libraries to participate in a shared digital library such as Internet Archive as much from the point of view of costs (free), functions (optical chapter recognition and conversion into EPUB and MOBI for e-readers directly implemented on archive.org) and permanent archiving (multiple mirror servers around the world) as from that

[1] The number of books that have been printed since Gutenberg's invention of the printing press is estimated at 129,864,880 by Leonid Taycher, an engineer at Google, according to an article published on his blog on August 5, 2010.

of visibility. Indeed, the position of a website in the list of Google search results depends on its PageRank. This depends largely on the number of links that point to its domain name. Under these conditions, a digital library with a large amount of content will automatically have a better PageRank and better visibility on the web and will therefore generate much more web traffic than a small digital library with very little content.

As Waibel [WAI 08] maintains, two schools of thought exist: an old school that believes that each library needs to create its own digital library and attempt to attract Internet users to it, and a new school that instead believes that in going beyond institutional communication and better satisfying the needs of Internet users, libraries would be better off participating in the digital libraries collectives already visited by Internet users, such as Internet Archive or even Flickr. This is also our point of view. With enough web traffic, libraries may prompt the participation of Internet users.

The introductory part of the book attempts to articulate its context and the methodology that was used.

Chapter 1 addresses the philosophical, political and economic representations of crowdsourcing and its consequences regarding the way in which libraries function. This conceptual chapter contains, in particular:

– a critical discussion regarding the definition of crowdsourcing;

– an original chronology of its historical origins;

– an analysis on the subject of its conceptual origins in philosophical currents that are sometimes diametrically opposed and, in particular, a conceptual contribution around the law of value;

– a reflection on the concept of the wisdom of crowds;

– an analysis of the diverse critiques of crowdsourcing applied to digital libraries that some people could today describe as the "uberization" of digital libraries.

Chapter 2 contains a selection of projects through types of tasks including:

– putting content online and participative curation;

– digitization and printing on demand in the form of crowdfunding;

– participative correction of OCR and participative transcription of manuscripts;

– folksonomy.

This chapter contains data and information collected from the literature for each project.

Original analyses for each major type of project are given in the conclusion of Chapter 2.

In Chapter 3, analyses from the point of view of information and communication sciences and a state of the art are offered with, notably:

– an original taxonomy of crowdsourcing in digital libraries distinguishing explicit (or conscious), voluntary and paid crowdsourcing and implicit (or unconscious) crowdsourcing, gamification and crowdfunding;

– an analysis of the motivations of libraries and the conditions necessary for the development of crowdsourcing projects;

– a taxonomy of the motivations of Internet users who contribute to their projects;

– analyses of the possible rewards and remuneration;

– clarification regarding the communication necessary for recruitment;

– developments in the specific community management of this type of project;

– analyses of the question of the quality and reintegration of the data produced;

– a reflection on the evaluation of crowdsourcing projects.

1

A Conceptual Introduction to the Concept of Crowdsourcing in Libraries: A New Paradigm?

1.1. A rapidly growing economic model

1.1.1. *What made this new economic model possible*

Internet users are growing more and more numerous and the time that they spend surfing the Internet is growing. The online encyclopedia Wikipedia required 100 million cumulative hours to be constructed. As Clay Shirky stated on August 28, 2008 at the *Wiki-Conference NYC*, if Americans, who watch 200 billion hours of television every year, used that time for creative activities instead, they could create 2,000 projects such as Wikipedia each year instead of watching television.

During a 2011 TED conference, Luis Von Ahn[1] claimed that using only 100,000 people, humanity succeeded in building pyramids and digging the Panama Canal, and that because of the Internet and social networks, it is now possible to assemble 750 million people, for example, for a project correcting the Optical Character Recognition (OCR) such as reCAPTCHA. An amazing "reservoir of goodwill" is therefore potentially available for cultural institutions if they know how to benefit from it.

[1] See: https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration (consulted June 23, 2016).

Digital Libraries and Crowdsourcing, First Edition. Mathieu Andro.

Participatory models came about with the development of the Web 2.0. The term was invented by DiNucci in 1999 [NGU 12] or by Dale Dougherty in 2004 [SAR 14] and popularized by Tim O'Reilly in 2005 [TRA 08]. Crowdsourcing now means that Internet users no longer have to be content with passively consuming Web content within a hierarchical, unilateral and static diffusion model (Web 1.0), but can actively participate in its development. The diffusion of information has become reciprocal, interactive and dynamic. The Internet user therefore ceases to be a consumer, a reader and a passive receptor who is content to browse, and becomes a producer, an author, an active emitter of information, a contributor who can participate in the writing and modification of content on the Web (comments, tags, wikis, social networks, etc.) and in the production of data and metadata. The authority of data has thus been moved from the server to the customer [BAI 12]. As telecommunications expert Benjamin Bayart emphasizes, if printing taught people to read, the Internet is now teaching them to write[2].

Well before Web 2.0, the invention of "self-service" which granted the consumer direct access to merchandise without the intermediary of a vendor and which was applied to libraries in the form of open access collections, was an early form of the integration of the consumer into the production process. This economic model was invented by Aristide Boucicaut in his department store "Le Bon Marché" whose slogan was "self-service, free to touch" giving customers, as described in Zola's *Au bonheur des dames* (translated into English as *The Ladies' Delight* or *The Ladies' Paradise*), the opportunity to access the merchandise actively and freely, without a shopkeeper as an intermediary, and, *in fine*, to take over part of the merchants' and store owners' jobs. Broadly speaking, production seems to have thus progressively lost the central place that it occupied in favor of consumption and the consumer society that developed after the Second World War.

Later, the "just in time" model, developed at Toyota, consisted of producing products "on demand" for the customer in order to avoid unsold stock by producing just-in-time supply in a way that is synchronized with and driven by demand. This model of "manufacturing without waste", "lean manufacturing" or "fat-free manufacturing" consists of producing only what you strictly need, with the necessary correct means, at the time when it is needed and at the least possible cost to the producer to externalize the

[2] See: http://www.gameblog.fr/blogs/poufy/p_58428_l-imprimerie-aura-permis-au-peuple-de-lire-internet-lui-a-pe (consulted June 23, 2016).

decision to begin production with the consumer. This model was born from the difficulty Japanese stores had in stocking merchandise due to insufficient space and the necessity of resupplying only when stock ran out. It was also significantly inspired by the way in which supermarkets operate. In the same way, the clothing chain Zara keeps only a single month worth of inventory and thus better adapts its production to trends in the market, producing models depending on sales [SUR 04]. Advertising itself participates in the integration of the consumer into the production process. Indeed, when we view a television program or website, we produce statistics and data, or when we view advertisements, we also produce value. We can therefore talk about an economy of attention [CIT 14]. The decision to visit this or that site could therefore be likened to a vote, a vote that participates in production and revenues of the producers. This model has found its application in libraries, in on-demand digitization by participatory financing (crowdfunding) and in printing on demand, which will be addressed in this book.

Today, crowdsourcing continues the relatively old movement of integrating the consumer into the production process. It was made possible by the development of the technologies of Web 2.0. Born from a cultural evolution toward more participative and collaborative approaches, crowdsourcing was made technologically possible by Web 2.0, that is to say, the possibility of having a large number of people, who have free time available on the Web, work remotely on collective projects. It is especially inspired by the way communities of freeware developers work. By calling on a crowd of Internet users, it is possible to carry out, in very little time, tasks that previously would have been impossible to complete or even imagine, or that would have required huge amounts of time. In short, crowdsourcing “is a way to find a needle in a haystack”, as Lebraty and Lobre [LEB 15] state. Sagot *et al.* [SAG 11] talk about “myriadization of divided work” and microworking. We could also talk about the “taskification” of work. Crowdsourcing has some similarities to the construction of medieval cathedrals, which required the capacity to “think big”, to delegate, to organize every task and above all to mobilize a large number of people around a common vision and goal, as Levi [LEV 14] recalls. It is also, to take a more recent example, what Alfred Sloan of General Motors described as “group management”, which consists of the solicitation of numerous collaborators to make the most important decisions.

We illustrate this idea with contemporary works of art in Figures 1.1 and 1.2.

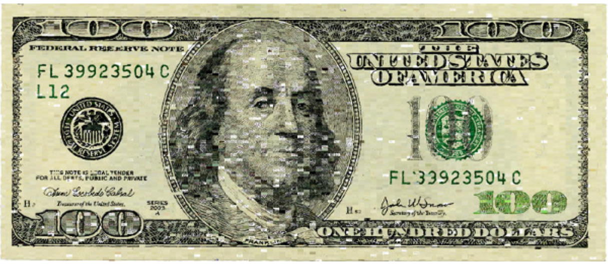


Figure 1.1. *The artwork Ten Thousand Cents*[3]*. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

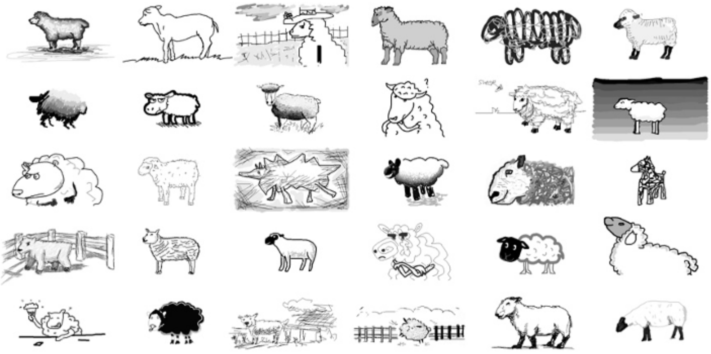


Figure 1.2. *An artwork juxtaposing sheep*[4]

In addition to art, crowdsourcing has already found applications in many areas. For example, in the field of video, YouTube and DailyMotion could not function without content posted online by Internet users. Crowdsourcing has also found applications in music, politics, fashion, banking, tourism, innovation, cartography, the search for missing planes, medicine, scientific research, publishing, translation and journalism. Using crowdsourcing is also topical in the field of GLAM (galleries, libraries, archives and museums) and digital libraries in particular, which is the subject of this book.

[3] This work of contemporary art created by Aaron Koblin was produced by 1,000 people working separately, using the Amazon Mechanical Turk Marketplace (AMT), on creating a milli-inch of a $100 bill without being aware of the purpose.

[4] These sheep were drawn by paid Internet users on the same AMT platform and were assembled by the artist Aaron Koblin (http://www.thesheepmarket.com).

1.1.2. *Application to digital libraries*

For libraries, digitizing and diffusing their collections on the Web means that they find themselves in the same space as their users. This situation makes possible multiple synergies and collaborations. Among cultural institutions, the amount of content that they make available on the Web has grown exponentially and there is no lack of painstaking work in indexing, describing and correcting this content. However, their budgets and their workforce have experienced an opposite trend which often leaves them sorely lacking. This state of affairs makes many goals impossible and the carrying out of other projects unimaginable without external aid. In addition, the real or virtual publics of these institutions are less and less content with the role of passive consumer of cultural information and would increasingly like to get involved in service to heritage and culture. In cultural institutions, the idea of being receptive to interaction with a participating public and volunteers largely preceded the emergence of the Web 2.0. However, the Relational Web has fostered the emergence of a participative culture on which the model of crowdsourcing in libraries feeds.

In digital libraries, crowdsourcing thus makes it possible to complete tasks that would be impossible to undertake without the help of volunteer Internet users, in the absence of financial and human means. This means, for example, to improve the quality of metadata or to enrich it (comments, tags, analyses, etc.), to benefit from the knowledge and skills of scholars, to develop communities around projects, to increase visits to the resources produced, to make the general public more aware of the conservation of common cultural heritage, to generate more interactions, innovative ideas and collaboration. For example, within the online public, there might be someone who would know how to identify a church in a photograph, a scholar could provide information about its construction and its history, an elderly villager able to identify a person in the photo, etc. The knowledge that teams of librarians have access to is much too limited to be able to respond to all of these questions. The knowledge present in the crowd of Internet users is limitless.

The British Museum understood this well when, on August 3, 2015, it published a call to Internet users on britishlibrary.typepad.co.uk with the title, “Help Us Decipher this Inscription”. Between August 3 and 18, 2015, the post had been shared almost 32, 000 times and had generated more than 11, 000 shares on Facebook and 9,000 tweets, as well as 115 comments directly on the blog between August 3 and 10.

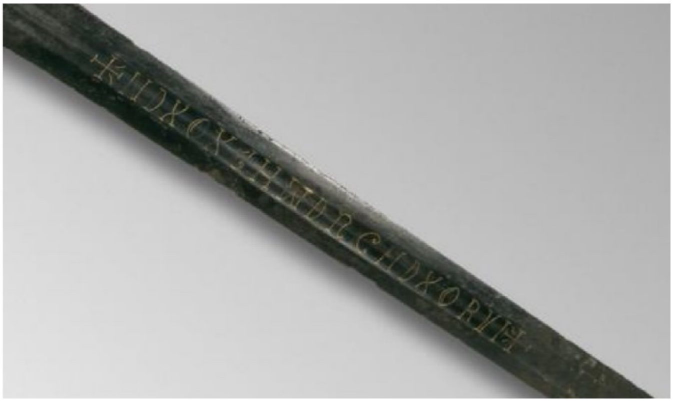


Figure 1.3. *13th Century sword whose photograph was published by the British Library[5]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

In order to mobilize Internet users, cultural institutions possess solid advantages. They often already have solid experience in mobilizing volunteers and organizing contests, reading groups and events and even in the "adoption" of books whose purchase is financed by readers or patrons. Furthermore, these institutions enjoy a positive image among the public and are considered to be trustworthy to work for the general interest and whose goals are cultural, not financial. These goals are therefore likely to attract volunteers and elicit contributions.

Crowdsourcing in the service of digital libraries is also the means of turning the sometimes thankless work required of a single employee into a worthwhile activity offered to an indefinite group of volunteer Internet users and "worker bees" who would like to actively contribute to the development of the cultural Web. The documents digitized and put online are thus the object of a participative redocumentarization, a remediation making it possible for new and collaborative processing of collections of documents by calling and sometimes on testimony and memory, and sometimes on the expertise and knowledge of Internet users. The collections are thus revisited, reinvented and reimagined.

[5] See: http://britishlibrary.typepad.co.uk/digitisedmanuscripts/2015/08/help-us-decipherthis-inscription.html (consulted June 23, 2016).

1.1.3. *Growing interest from politicians, Internet users and academics*

The success of crowdsourcing projects and the interest in these projects from Internet users, politicians and academic researchers, is increasing. As Sarrouy [SAR 14] reports, a 2011 study by massolutions.com estimated the crowdsourcing market at more than 300 million dollars with a growth rate of more than 75% between 2010 and 2011. In 2012, another study by [MCK 14] evaluated the gains in productivity, calculating social media and crowdsourcing platforms in consumer goods, financial services, advanced production and professional services at 25%. Finally, at the end of 2013, the Gartner firm anticipated that by 2017, more than half of producers of consumer goods will base more than 75% of their research and development on crowdsourcing. In the area of citizen science involving biodiversity alone, researchers at the University of Washington estimate that the in-kind contributions of the 1.3–2.3 million volunteers would have an economic value of more than 2.5 billion dollars.

Crowdfunding, in particular, would have been able to finance a million projects in 2012 and raise 2 billion euros [ONN 13]. Although the financing of projects by private individuals in itself is nothing new, the Internet makes it easier to do and to gives a new scope to participatory financing that already represents a market of three billion dollars worldwide in 2012 and whose growth is exponential.

By using the service Google Trends, which is to say the traces left involuntarily[6] by Internet users who perform Google searches, we also observe that, beyond politics, more and more Internet users entered the word “crowdsourcing”, which has very few translations into modern languages, into the Google search engine starting in 2006, when the term was popularized by Jeff Howe. In a base 100 system, the countries whose Internet users carried out the most searches containing the word crowdsourcing are in order as follows: the Netherlands (100), Portugal (60), Germany (60), Spain (56), Singapore (55), Austria (54), Switzerland (54), the United States (48), Brazil (43) Denmark (38) and the United Kingdom (31).

[6] In this case, we can talk about “implicit crowdsourcing”, which means involuntary contribution, as we will see later.

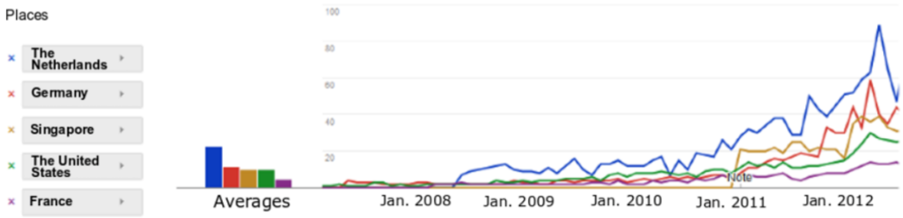
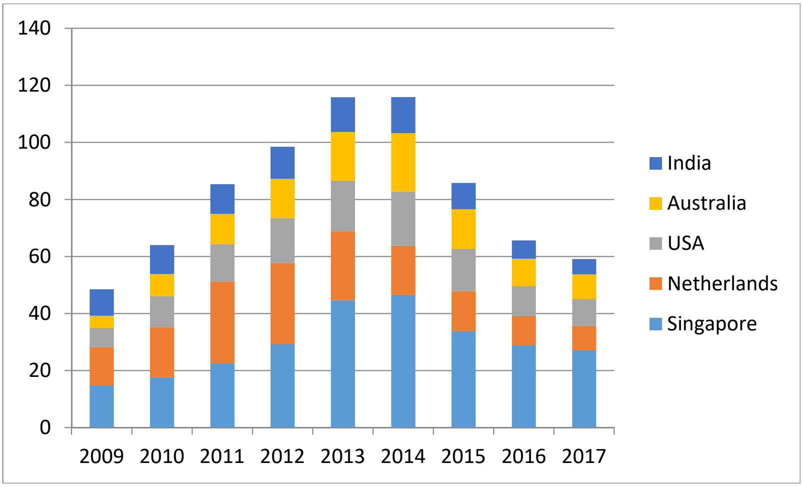


Figure 1.4. *Change in the number or searches for the word "crowdsourcing" on Google for each country, according to Google Trends. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*



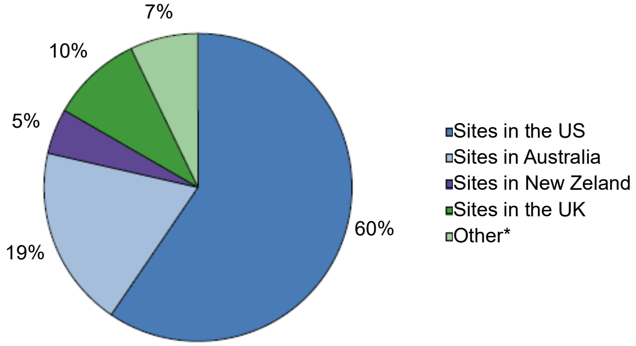


Figure 1.5. *Countries represented in the survey conducted by OCLC about social metadata, from [SMI 11]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

An investigation of crowdsourcing projects applied to libraries was carried out by the OCLC [SMI 11]; it showed that, among the projects studied whose leaders were sought for the investigation, 60% were American, 19% Australian, 10% English and 5% New Zealander, and only 7% were from other countries of the world.

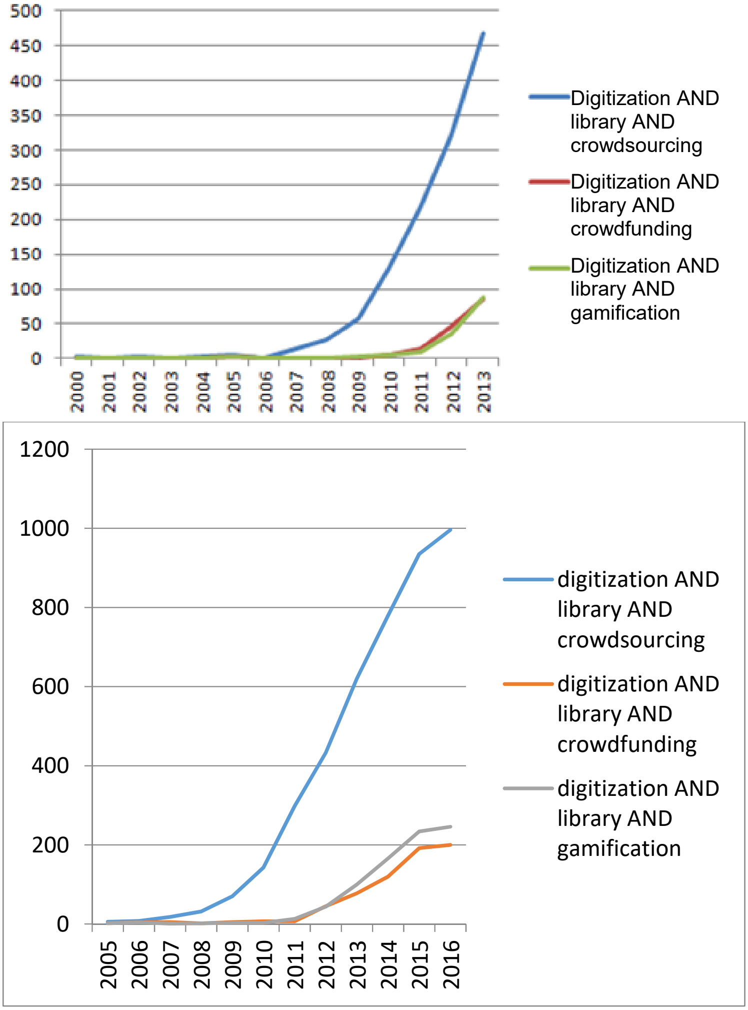


Figure 1.6. *Change in the number of publications on crowdsourcing indexed by Google Scholar applied to the digitization of libraries. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Crowdsourcing applied to digitization projects therefore should not be considered a purely Anglo-Saxon phenomenon.

The interest of scientific research worldwide in the phenomenon of crowdsourcing is growing, especially with regard to its applications to digitization of the heritage preserved in libraries. This statement can be supported by observing, for example, the number of articles indexed in Google Scholar about this specific subject.

1.2. Origin, definition and scope of crowdsourcing

Crowdsourcing has long been a pragmatic professional practice well before it was conceptualized and became a subject of academic research. Under these conditions, its origin, definition and scope can be difficult to establish. Before becoming a buzzword, the term "crowdsourcing" was first used by Jeff Howe in the title of an article published in *Wired Magazine* in June 2006, which was entitled "The Rise of Crowdsourcing". According to [SCH 10], the term had, however, been used by an anonymous Internet user in a forum. Other authors prefer to talk about "open work" or "fair-trade work".

In the case of digital library projects whose actual contributors are only an active minority of volunteers and cannot, in any case, be assimilated into a crowd, certain authors prefer to use the term niche sourcing or community sourcing, preferring the more specific word "community" to that of a more indeterminate "crowd". It involves not so much using the public than recruiting volunteers motivated by a spirit of collaboration, cocreation and co-construction. This idea is related to the one laid out by Jakob Nielsen[7], according to whom 80% of Internet users are passive consumers and 20% are active contributors and producers of content on the Web. According to Holly Goodier[8], these proportions would have changed since then and would now more likely be 25% of Internet users who are inactive, 45% who comment and enrich and 30% who produce content. When it comes to digital libraries, the term *community sourcing* seems the most judicious to us. We nevertheless will use the term crowdsourcing, which is more common, will make our

[7] See: https://www.nngroup.com/articles/community-is-dead-long-live-mega-collaboration (consulted June 23, 2016).

[8] See: www.bbc.co.uk/blogs/bbcinternet/2012/05/bbc_online_briefing_spring_2011.html (consulted June 23, 2016).

writing more intelligible and will allow us to avoid resorting to complex jargon.

The authors of [EST 12], whose work is authoritative, have sought to work specifically on the question of the definition of crowdsourcing by collecting, in the literature, the diversity of definitions that are found there. No less than 40 citations in 32 articles published between 2006 and 2011 were collected in this study that has categorized the different elements necessary for the construction of a summary definition.

Who makes up the crowd?	Amateurs.
What does the crowd do?	It voluntarily and consciously accomplishes tasks and microtasks in order to solve problems.
What does the crowd get in return?	Distraction, pleasure, the development of skills, experiences, knowledge, the sharing of knowledge, the love of a community, economic compensation, social recognition or better self-esteem.
Who initiates it?	Public or private companies.
What type of process is involved?	A production process, an economic model, participative outsourcing of a task after a request that is open to everyone.
What medium is used?	The Internet.

Table 1.1. *Multicriteria definitions of crowdsourcing*

Based on these elements, here is the definition which these authors have come up with:

> "Is a type of participative online activity in which an individual, an institution, a non-profit organization, or company 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, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always 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". [EST 12]

The question of the voluntary or involuntary nature of the participation of Internet users can nevertheless be discussed. Indeed, if we believe that the contribution is necessarily voluntary as this definition asserts, we exclude the field of crowdsourcing on sites such as YouTube, OCR correction resulting from reCAPTCHA and a large part of the projects that collect contributions of Internet users in the form of games (gamification). If we recognize that this contribution is not necessarily voluntary, the scope definitely expands considerably. In every case, excluding not fully conscious forms of participation from the field would at least deserve justification, which seems difficult. Maybe it is therefore preferable, from our point of view, to speak rather of explicit crowdsourcing when the contribution of Internet users is voluntary and implicit crowdsourcing (or involuntary crowdsourcing or passive crowdsourcing) when it is not [HAR 13]. Renault [REN 14b] also considers this definition to be somewhat naive, since there are many contributors to crowdsourcing who are not aware of their contribution. Nevertheless, one could consider implicit crowdsourcing as a sort of betrayal of crowdsourcing, which was initially conceived as a means of rehumanizing the Web, and see it as revenge of the commercial Web on the power of Internet users. Indeed, with implicit crowdsourcing there is a large risk of taking advantage of citizens for the benefit of lobbies, to consider Internet users and the traces that they leave on the Web, especially with their connected devices, as simple means without connecting them to projects [LEC 13].

Schenk and Guittard [SCH 12] has also made the choice to place this form of crowdsourcing in its typology by describing it as "non-voluntary" and by establishing a parallel with the concept of positive externality. Implicit crowdsourcing could, in fact, be considered in light of the concept of positive externality (or external economy). In this way, by the traces that they leave or by their unconscious work, Internet users, as economic agents, provide an economic service that can be exploited for other agents without being compensated. Thus, Google benefits from the work of Internet users who unknowingly correct its OCRized texts by reentering reCAPTCHAs in order to prove that they are not robots so that they can create accounts on websites: just as a beekeeper benefits implicitly from the work of an arborist since the former's bees can gather pollen from the flowers on the trees that the latter cultivates, without financial compensation, in return, the bees will also support the fertilization of the trees [MEA 52]. In the case of Google Books, the company could indeed thank its involuntary contributors or be taxed for this hidden work. However, one could equally estimate that the improvement by Internet users of the quality of the texts accessible to those Internet users for free, benefits them directly in return.

Taking all of these considerations into account, crowdsourcing can therefore be defined, after reading a representative group of publications and according to the definition that we present in Box 1.1.

Crowdsourcing is a form of outsourcing that allows the contribution of work, money (crowdfunding), skills, knowledge, intelligence, creativity or experience, through voluntary (explicit crowdsourcing) or involuntary (implicit crowdsourcing) engagement of Internet users. This outsourcing is carried out following an appeal to an individual, an institution or an organization. Internet users will gain, in exchange for their contribution, social recognition, experience, the acquisition of skills, compensation or remuneration (paid crowdsourcing). They can also act to improve self-esteem through distraction, pleasure, love for a community or disinterested altruism.

Box 1.1. *Definition of crowdsourcing*

Now that this definition has been introduced, in order to fully understand what crowdsourcing is, it seems necessary to define its scope by laying out what crowdsourcing is not. Indeed, the concept of crowdsourcing is somewhat close, for example, to that of *human computation* that evokes the possibility of having humans and their collective intelligence do tasks that computer programs are still incapable of carrying out automatically. Nevertheless, crowdsourcing is distinguished by its simpler and less sophisticated tools and tasks and by contribution rules constructed in a more collaborative way.

With crowdsourcing, the strength of the crowd resides more in the aggregate of independent ideas than in their collaboration [SZO 12]. It is therefore also distinct from collective intelligence.

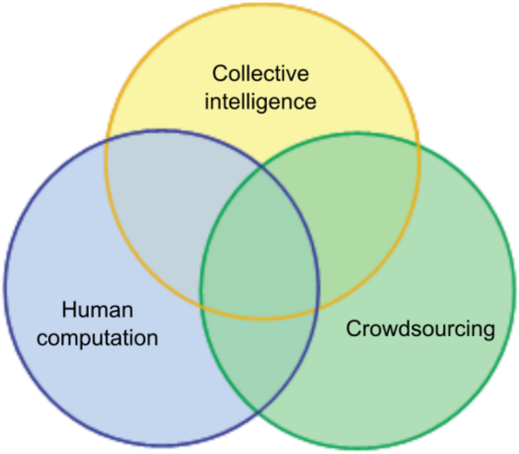


Figure 1.7. *Relationships between human computation, collective intelligence and crowdsourcing, according to [HAR 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

User innovation is a form of opening up research to Internet users. It is much more open than crowdsourcing, which is finally very circumscribed by the contribution process. It consists of collecting research ideas and innovations of Internet users, most of the time in the form of contests and calls for contributions that generally lead to compensation. The history of science is, in fact, full of innovations coming from amateurs outside the profession who are hobbyists and who, not seeking to reproduce established models with which professionals were trained, are sometimes likely to lead to innovative ruptures. MIT researcher Von Hippel, who talks about innovations through use or bottom-up innovations, estimates that 46% of companies in the United States in innovative sectors have their origins in a user. Innovation has become, because of their contribution, the result of a direct collaboration between the producers and consumers who become coproducers. In science, the phenomenon of “unexpected readers”, accidental discoveries and happy coincidences (serendipity) are well known and are a good example of this phenomenon. However, crowdsourcing is also distinct from the logic of user innovations since in the latter case, the business is not always the initiator and origin of the projects and ideas from which it benefits via the suggestions of consumers. With crowdsourcing, the business remains the initiator of the projects.

Crowdsourcing is also different from *open innovation*, since unlike the latter, it is a form of outsourcing to the crowd of Internet users via Web 2.0 and not the outsourcing of innovation to other companies.

The concept of outsourcing nevertheless corresponds to that of crowdsourcing, since the approach resembles the one used within the framework of an open tender with the publicity that the request is given. It involves outsourcing certain missions not to a specific service provider, but to an undefined community of volunteer Internet users in order to be able to carry out projects or innovations that would have been impossible without them. Crowdsourcing could thus be considered simultaneously as a revised form of outsourcing, an innovative economic model and an alternative to subcontracting. However, unlike outsourcing, crowdsourcing does not require a contract between the sponsor and the service provided, as much as it involves a large and undefined number of collaborators.

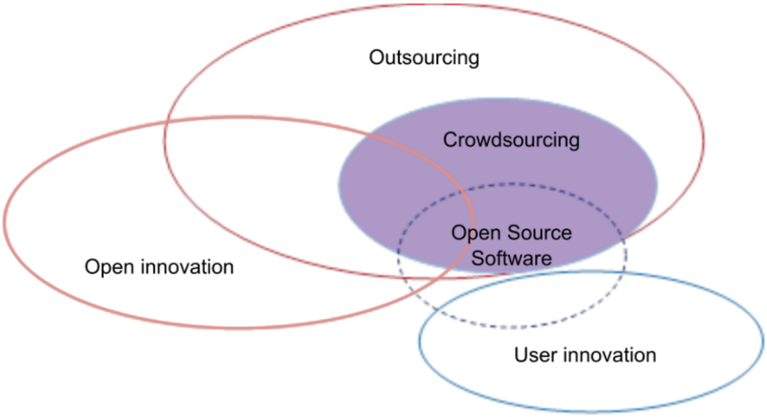


Figure 1.8. *Position of crowdsourcing among neighboring areas, according to [SCH 10]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Finally, crowdsourcing could be considered the application of Open Source methods from other industries outside of software. Nevertheless, developments are not always made in an exclusively collaborative way and can also be fed by the spirit of competition. Moreover, while Open Source is based on several contributors working to satisfy the needs of several users, crowdsourcing is based on the idea that several contributors will work in the service of a single entity.

We have distinguished five large families of crowdsourcing projects applied to digital libraries and we have offered an original taxonomy containing explicit crowdsourcing, implicit crowdsourcing, gamification, paid crowdsourcing and crowdfunding.

1.2.1. *Explicit crowdsourcing: using volunteers*

If traditional explicit crowdsourcing shows the interest in collaborating with the general public and the company and in the source of opportunities through the disruptive innovations that the public can sometimes create, the market still available for this revisited use of volunteering is, nevertheless, beginning to tighten because of the multiplication of projects and the appearance of new forms of crowdsourcing. Furthermore, the benefits drawn from these projects do not always compensate for the significant investments necessary for the development of platforms, communication, recruiting, training and management of communities of volunteers.

1.2.2. *Implicit crowdsourcing: using involuntary and unconscious work*

Implicit crowdsourcing consists of having Internet users work without their being aware of it. This form of crowdsourcing has made it possible to obtain excellent results, but can pose ethical questions.

1.2.3. *Gamification: using players*

These projects, which consist of obtaining work from Internet users by having them play, can be expensive to develop and can also obtain excellent results, but the collaboration is essentially smaller with Internet users, who sometimes benefit less when it comes to personal development.

1.2.4. *Paid crowdsourcing: using microemployees*

This form of crowdsourcing popularized by the Amazon Mechanical Turk Marketplace and widely used in the United States has sometimes been criticized as a form of exploitation of work outside of any regulatory framework. However, by using this type of crowdsourcing, libraries are also making the choice to use their budgets to benefit contributors rather than development of platforms and communications campaigns for recruiting. The Amazon marketplace has already been developed and connects public or private businesses that offer microtasks (classification, indexing, identification, transcription, correction, editing) with more than 700,000 workers already recruited from around the world and at a price that they fix voluntarily.

1.2.5. *Crowdfunding: institutional "begging"*

This form of crowdsourcing does not employ the work of volunteers, but instead uses their money. It has already been used successfully to finance projects. Participatory financing (or micropatronage or patronage on demand) is a specific form of crowdsourcing to which the contribution of Internet users is exclusively financial.

Beyond this introductory section meant to define crowdsourcing in order to better define the boundaries of this book, we will revisit the definition of crowdsourcing more thoroughly by applying it specifically to the domain of digital libraries which interests us and by producing a more detailed original taxonomy of crowdsourcing in digital libraries. These developments will find their place in Chapter 3, which is dedicated to analyses from the perspective of information and communication sciences.

1.3. Historical chronology of crowdsourcing

Crowdsourcing could be said to date back to Hugues de Saint-Cher, a Dominican in the 13th Century who coordinated numerous monks in order to index the content of holy texts [LED 15].

However, the majority of authors date the beginning of the history of crowdsourcing to the "Longitude Act" of 1714. After the accident of the English admiral Cloudesley Shovell in 1707 in the Isles of Scilly, the government decided to offer 20,000 pounds to anyone capable of determining the longitude of a ship on the open sea and avoid more accidents [DAW 11]. The famous scientists Cassini, Huygens, Halley and Newton were unable to find a solution and it was John Harrison, a carpenter and watchmaker, who won the prize from among more than a hundred competitors [LAK 13].

In 1726, an order from Louis XV required ship's captains to bring back plants and seeds from the foreign countries that they visited [BOE 12] and thus contribute to botanical research.

Several decades later, in 1758, mathematician Alexis Clairaut was able to calculate the orbit of Halley's comet by dividing the calculations tasks between three astronomers. For his part, British astronomer Nevil Maskelyne calculated, in 1750, the position of the moon for navigation at sea because of the comparison of the calculations of two astronomers who carried out the calculations two times each, which were then verified by a third party.

In 1775, Louis XVI offered a reward to whomever would make it possible to optimize the production of alkali, a chemical product. The competition was won by Nicolas Leblanc [CHA 15].

In 1794, French engineer Gaspard de Prony organized microtasks of addition and subtraction for 80 unemployed hairdressers in order to develop detailed logarithmic and trigonometric tables.

In 1850, 600 volunteers in North and South America sent meteorological data to scientists at the Smithsonian Institution using telegraphs [STE 14].

In 1852, the deparment store "Le Bon Marché", founded by Aristide Boucicaut, offered a self-service store for the first time, ancestor of today's supermarkets. Part of the producer's work is thus externalized to the consumer. The self-service model would find other applications in commerce (automatic cashiers, for example) and applications in banks (cash dispensers), hospitality (in fast food restaurants, for example, consumers are the ones who provide the service and clear the table), interior furnishings (consumers are the ones who assemble the pieces of IKEA furniture, for example), transportation, laundromats for clothing or vehicles and libraries (open access collections).

In 1857, the *Oxford English Dictionary* benefitted, following a call for volunteer contributions, from more than 6 million documents containing proposals for words and citations of use.

In 1884, the Statue of Liberty was financed following a public subscription of 125,000 people which had been started in France in 1875.

In 1893, Francis Galton, an English statistician and the father of eugenics, observed, during a competition launched at a livestock market which involved guessing the weight of a steer, that the average estimate of the crowd was closer to the truth than the estimate of experts, implying the existence of a wisdom of crowds.

In 1894, librarian James Duff Brown allowed readers at the Clerkenwell Public Library direct access to part of its collections. Open access in libraries was born; it is the adaptation of the self-service model to libraries.

In 19th-Century France, the government sent out calls for contributions. One of them, won by Nicolas Appert, allowed for the discovery of new methods for conserving food in the form of canning.

In the 19th Century, in the field of publishing, the public subscription system was developed to finance the publication of books.

In 1900, the National Audubon Society (United States and Canada) organized an annual bird count, the “Christmas Bird Count”.

In 1936, Toyota assembled 27,000 people and selected one design to become the brand’s logo. Much later, the logos of Nike and Twitter, for example, would be directly inspired by consumers.

In 1938, in the United States, the Mathematical Tables Project mobilized 450 unemployed people, victims of the Great Depression led by a group of mathematicians and physicians, in order to calculate tables of mathematical functions, well before the invention of the computer.

In the 1950s, an industrial engineer at Toyota, Taiichi Ōno, invented the “just-in-time” model, ancestor of the “on-demand” model, which made it possible to produce, without stock or unsold goods, with a lean supply chain according to demand. It involved, in a sense, outsourcing the decision to produce to the consumer. This model is, in the field of libraries, the origin of digitization on demand by crowdfunding and of printing on demand.

In 1954, the first telethon in the United States was able to collect funds to fight against cerebral palsy.

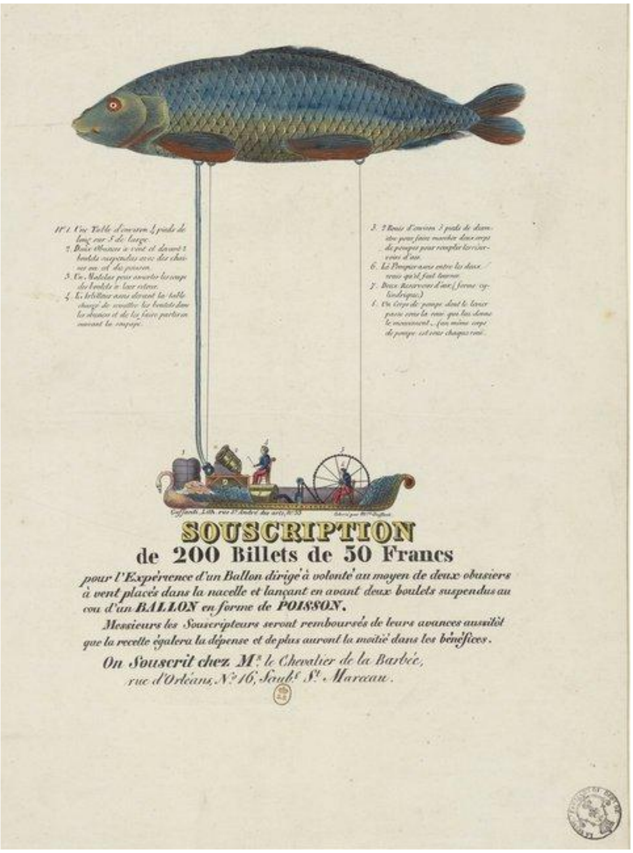
In 1955, the Sydney Opera House was designed and built following a public competition that encouraged ordinary people in 32 countries to contribute to the design project.

In 1979, the Zagat Survey, a restaurant guide, based its reviews on a large group of testers. The project was bought by Google in September 2011.

In 1981, the travel guide *Lonely Planet* was written, for its third edition, in a participatory way by independent travelers.

In 1997, the rock group Marillion financed a tour in the United States using donations from its fans totaling $60,000.

In 1998, the directory Dmoz offered content generated by its users. The Web 2.0 was born.



SOUSCRIPTION

de 200 Billets de 50 Francs

pour l'Expérience d'un Ballon dirigé à volonté au moyen de deux obusiers à vent placés dans la nacelle et lançant en avant deux boulets suspendus au cou d'un **BALLON** *en forme de* **POISSON.**

Messieurs les Souscripteurs seront remboursés de leurs avances aussitôt que la recette égalera la dépense et de plus auront la moitié dans les bénéfices.

On Souscrit chez M.r le Chevalier de la Barbée,
rue d'Orléans, N.o 16, Faub.g S.t Marceau.

Figure 1.9. The first form of crowdfunding.
From http://gallica.bnf.fr/ark:/12148/btv1b8509563b (consulted June 23, 2016).
For a color version of the figure, see www.iste.co.uk/andro/libraries.zip

In 2000, the philanthropic crowdfunding platform justgiving.com appeared, along with the participative financing platform artistshare.com which would be followed by multiple initiatives to this day.

On November 23, 2013, the video game *Star Citizen* collected an amount of $30,044,586.

At the end of 2005, Amazon launched the crowdsourcing platform Amazon Mechanical Turk Marketplace, making it possible to connect businesses and institutions searching for workers on the Web around microtasks.

1.4. Philosophical and political controversies

Crowdsourcing is a subject that can be the source of strong ideological divides. In the sections that follow, we have confined ourselves to reporting, in the most balanced way possible, the analyses and advantages and disadvantages pushed to the fore by this or that theorist or by this or that ideology.

The philosophical and political origin of crowdsourcing can seem very confused at first glance. This economic model seems, in fact, to be able to echo ideologies as diametrically opposed as Marxism and liberalism. However, at the end of this chapter we will see that a certain coherent synthesis between these opposites can be delineated by means of "Californian ideology".

There seems to be a relationship between crowdsourcing and socialist ideologies. Is not it for this reason, accused citizen science of Lysenkoism[9], of being affiliated with "proletarian science" and of representing a desire for popular control of science, of representing an "attempt at ideological intrusion and the taking over of part of scientific output by ideological lobbies"[10]?

The Internet users who participate in crowdsourcing projects seem, in fact, to embody the socialist motto "from each according to his or her ability,

[9] Named after Trofim Lyssenko, a Russian agronomist of the 1930s who tried to apply Marxism to the natural sciences, the concept of "Lysenkoism" is usually used to refer to the intrusion of ideology into scientific research.

[10] Blog post "la faucille et le labo", reported by Lipinski (2014).

to each according to his needs”. Indeed, each one does his or her best to contribute to producing content according to the time, strengths and skills available to them. And the content produced will benefit everyone, those who really need it, the same as the others, and those who have contributed greatly, the same as the others. There is no proportion between what has been produced and what will be consumed. The law of value is bypassed.

Among the motivations of contributors to many crowdsourcing projects, we see the desire to sacrifice their time for the common good, the need to feel useful to a community, acting from altruism and accountability to protect cultural heritage, etc.

Certain authors, such as Jean-Pierre Gaudard in his book *La fin du salariat*, herald the disappearance of the wage-earning class. With Generation Y’s arrival in the job market and in particular the development of freelance or miroentrepreneurial work, the relationship with work appears to be evolving. Engagement with the business seems to be weakening with the emergence of more workers who are more autonomous, individualist and more centered on the ego. The tension between the individual employee and the collective business seems to be increasing with the arrival of Generation Y on the job market. Digital natives are no longer invested in this collective framework; they are without attachments and no longer settle down. Often considered lawless mercenaries, they are sometimes also homeless, searching for a lost identity and suffering from a lack of recognition and difficulty finding fulfillment within the confines of a traditional business. At the same time, a “creative class” seems to be emerging. Therefore, we talk about jobcrafting, i.e. the process in which employees actively and gradually revise their job descriptions and their relationships with others [DEN 13]. The concept of work tends to disappear to the benefit of the concept of activity.

With crowdsourcing, if consumption becomes a producer of value and leisure becomes a creator of wealth, then work becomes a leisure activity. Money seems to no longer be the principal motivation of a significant crowd of amateurs who, in the positive sense, gravitate toward the domain of freeware to the detriment of motivation, self-investment and passion. On the Web, economic models based on being free also appear to predominate, suggesting the emergence of “collaborative commons” [RIF 14], of a “contributory economy” [STI 15], a sharing economy, a “participatory economy” or a “collaborative economy”. According to [STI 15], with automatization making work less and less necessary, employees, like consumers, will become contributors to the business, in other words, amateurs motivated more by their centers of interests than by their economic interests. It then becomes a question of paying them according to contributory profit sharing. Already, some businesses no longer have employees, but use external contributors or workers via Amazon Mechanical Turk Marketplace. The Internet therefore seems to be the medium for the abolition of mediation. Therefore, there are numerous websites have contested their place as intermediaries between the producer and the consumer and traditional economic actors who are comfortably established or who enjoy monopolies (taxis, rental agencies, employment agencies, etc.). We even talk about an "uberization" of the economy. More and more businesses are thus risking being supplanted by web companies with access to more competitive self-employed workers. This movement is far from being marginal. Thus, according to a PwC study published in 2014 under the title *The Sharing Economy*, the collaborative economy should go from 15 billion in 2014 to 335 billion euros in 2025.

Some peer-to-peer (P2P) theorists, such as Michel Bauwens, are of the opinion that humans can now contact each other, share data and collaborate without permission or hierarchy, each one filling in the other's gaps and that this will profoundly change our societies. According to them, P2P is therefore the socialism of the 19th Century. Vertical hierarchies were defined by power. With P2P communities, it is reputation that predominates; they function in a more horizontal manner. This reputation is measured depending on web traffic generated by the production of a particular person on the same model as the number of citations in scientific research. We can even talk about the economy of reputation insofar as reputation can be converted into money via advertisements that pay according to the web traffic generated, but also in jobs and in opportunities for partnerships.

In any event, even if it turns out to be less revolutionary than certain theorists claim, crowdsourcing constitutes "a disruptive innovation, which will therefore profoundly and permanently change the business ecosystem" [LEB 15].

By seeking to rehumanize the Internet and by restoring the central place to the human as origin and purpose of a website which must be created by humans and for humans, crowdsourcing is also unquestionably a descendant of humanist philosophers and eudemonists. Crowdsourcing uses human crowds whose capacities and intelligence remain largely superior to those of algorithms. Faced with artificial intelligence and Big Data, crowdsourcing retains faith in human superiority. Moreover, the paid crowdsourcing project Amazon Mechanical Turk Marketplace mischievously has as its logo a very old automated chess player, which was said to have real artificial intelligence while, in reality, there was a person hidden in the mechanism. In this way, Amazon affirms that human intelligence remains unsurpassable.

Crowdsourcing is also part of the digital humanities movement; we can even talk, justifiably, about "digital humanism" following the example of Milad Doueihi, insofar as new technologies have a universal dimension and that they are the culture, since they set up a new context.

Crowdsourcing can just as well be considered a liberal, new and expansive form of outsourcing and opening of an organization to its outside environment. Indeed, in the first instance, globalization of the economy and heightened competition between businesses led industries, not recognizing any other law than that of supply and demand, to outsource to countries with low-cost labor. However, with the development of the Internet, it has now become possible to employ anyone and simply link them to the network. Crowdsourcing thus remains a form of outsourcing work on the Internet, in areas that are still limited.

On the Internet, links, clicks, comments, ratings, recommendations, visits links, etc., function like votes in a democracy. The sites that are well referenced and showcased by search engines are the sites elected by Internet users. There is a hierarchy between them since the most visible pages are the most cited, most linked to, the most commented on. PageRank could, in a sense, be considered a form of implicit crowdvoting [REN 14]. By adding, on the Web, a link to a website, the Internet user will thus unconsciously vote for the site to be better referenced by the search engine.

Crowdsourcing also extensively relies on the concept of the wisdom of crowds that is, itself, very close to the liberal concept of the invisible hand. Francis Galton, father of eugenics and cousin of Charles Darwin, noted that during a popular contest consisting of guessing the weight of a steer, the average of the participants' estimates was very close to the truth. Today we can observe, in the same way, that if we ask a lecture hall to guess the number of marbles in a bottle or the temperature of a room, the truth is very close to the average of the responses. It is for this same reason that participants in the gameshow *Who Wants to be a Millionaire?* had a much greater chance of getting the correct answer by soliciting public opinion than by asking a friend. Drawing on this phenomenon, the Intelligence Advanced Research Projects Activity (IARPA), an American intelligence agency was launched the Good Judgement Project in order to draw benefits from the wisdom of crowds, since they are likely to better predict geopolitical events than the experts and analysts traditionally used by intelligence agencies. This project is an echo, in a way, of the adage *vox populi vox dei* and the following quote from Machiavelli, who believed that "there is a good reason that people say that the voice of the people is the voice of God. We see public opinion forecast events in such a marvelous way that we would think that the people are gifted with the occult ability of predicting and fortune and misfortune. As for the manner of judging, we rarely see them be wrong" [MAC 37].

Once again in the area of intelligence, analysis by text mining of the geographical locations co-occurring the most with the name Bin Laden showed that those places were the closest to the place where he was actually found. This does not mean that journalists knew Bin Laden's location, this means that a large amount of data can be transformed into high-quality intelligence and that where there are crowds, there is science.

From this point of view, it would therefore seem clear that there is an "invisible hand" which allows freely associated individuals to find fair and harmonious solutions without the intervention of any kind of authority, and that unimpeded private interests would be naturally beneficial to the common interest. This notion is also close to that of spontaneous order, proposed by Friedrich Hayek, namely a self-generated, self-organized order without a plan or authority, like the one that rules over the markets, but also Holacracy, a fractal organization of organically self-organized teams, or sociocracy. One could consider the participative encyclopedia Wikipedia as another spontaneous order, since it is exhaustive and structured through the autonomous and uncoordinated action of individuals, without a complete plan existing before its development. Jimmy Wales, the founder of Wikipedia, moreover cites Friedrich Hayek, in particular for his conception of the Wikipedia project. In fact, the belief in the spontaneous correction of Wikipedia articles is somewhat similar to the liberal belief in the invisible hand of the market.

Organizations that use crowdsourcing are aware of their limits. They have confidence in the capacity of crowds to spontaneously find the best solutions when they return the freedom of initiative and autonomy to the individuals who make it up.

With the development of the new economy, the difference between public and private life, volunteering and work, seems to become more confused. Employees are working more and more on transportation, at night, during the weekend, on their vacations, etc. Conversely, they also sometimes dedicate working hours to social relationships, or even to leisure with the blessing of businesses that understand that their personal fulfillment will be a source of creativity and innovation. We sometimes talk about *weisure*, using Dalton Conley's expression, a mixture of work and leisure, or *playbor* or *playbour*, a mixture of play and labor, or "intrapreneurs", that is to say people who have the spirit of initiative and enterprise but are still employees, entrepreneurs inside the business. Hierarchies have been overturned, and it is no longer management who decide and employees who act, but often the employees who are directly responsible for projects. Open innovation calls into question the social division of work [VON 05]. With Web 2.0 and particularly with crowdsourcing, the border between producers and consumers is in the process of disappearing, since consumers of information on the Web are also becoming its producers. Millions of people produce data, for pleasure, and as a result work for free for YouTube or Facebook. Others participate in the improvement of software without knowing it when they use it for free. While Facebook announced a total revenue of 2.5 billion dollars in 2013, which equals $6.81 per active user, this revenue remained above all tied to advertising and not to the resale of data. When Internet users type a search request into Google, write a tweet, add content to Facebook, write a comment about a book on Amazon, post an evaluation of an eBay seller, review the quality of a restaurant on the Internet, they produce data that have value, which will be resold by these companies, and work for them for free in exchange for the free service that the company provides for them. Fuchs [FUC 12] estimates that in this way Facebook has benefitted from 60 billion hours of unpaid labor. On the Web, people use many applications that appear to be free. In reality, in exchange for the service being free, users work to produce data without even being aware of it: when they write on Facebook, copy a CAPTCHA and even perform a search. Data production work is free from any regulation or legislation.

Therefore, instead of the participation of Internet users, crowdsourcing could so instead lead to the exploitation of the free work of users sometimes referred to as *servuction*. Thus, Petersen [PET 08] reports, in 1999, seven of the 13,000 AOL volunteers who worked for free to sustain and energize the AOL community finally received payment for their work. Later, two of them even went so far as to submit a complaint against AOL to a federal court in New York before the inquiry was closed in 2001.

This method of working, which changes the borders between production and consumption, has been conceptualized under the term *digital labor*. It includes the implicit and invisible work of the production of data by Internet users resulting from their activities on the Web and exceeds its limits [CAR 16].

Be that as it may, in the face of the influence of certain sites earning their profits because of the unpaid work of Internet users, governments sometimes show a desire to develop a tax system around data capture. Subjecting data to taxes would make it possible to give the community a portion of the creation that it has provided in the form of "invisible work". But this work is all the more invisible because it is low intensity and difficult to recognize.

Digital labor could also be compensated in the form of individual micropayments, or in exchange for shares, in particular for crowdfunding (equity crowdfunding), or via collective taxation of data. With crowdfunding 2.0, the participants may therefore go from consumers to shareholders and start-ups sell stocks to finance their projects. A text along these lines was passed this way in the United States by the Securities and Exchange Commission (SEC). As explained on the blog InternetActu.net[11] in particular, the user could thus be recognized as a producer of data to regain control and be paid as a producer of value.

With crowdsourcing, we could go from a mode of production in which the proletariat sells its labor to the capitalist in exchange for a salary, to a participative economy in which the contributor offers his or her participation in the interests of a community of Internet users. The Amazon Mechanical Turk Marketplace, for example, following the example of other paid crowdsourcing platforms, allows an extension of freelance independent work, a new form of work: employers offering tasks on the platform and workers freely carrying them out as microentrepreneurs and outside of any rule other than the law of supply and demand in a totally open and liberal market where people freely sell and buy work from each other online. Instead of risking burnout in its employees, the employer can use this method to, in a few minutes, recruit

[11] See: http://www.internetactu.net/2012/06/01/vers-un-nouveau-monde-de-donnees (consulted June 23, 2016).

crowds of workers with diverse profiles who are available all the time, usually inexpensive, accessible without other administrative steps and paid only once the work is accomplished. The employer can thus carry out tasks that were impossible to imagine before. It can, in a few minutes, recruit a workforce that is just as large and diverse as those of large businesses and mobilize them around projects.

From the point of view of workers, some are happy to be able to work when they want, when they need to, as much as they need to and for whomever they like, and to choose the activities that they will do. Others make a living, for example, on the services that they provide through Uber as drivers, or doing odd jobs or gardening on TaskRabbit. They share the goods which they own, but of which they have limited use and are focused on quality and use rather than ownership in order to lower their expenses and avoid waste due to collaborative consumption [PEU 15].

However, from an ethical point of view, the exploitation of volunteer or underpaid work and the freedom from any legislation within the framework of Amazon Mechanical Turk poses a problem that is simultaneously legal, social and even economic. It should be noted that it also involves, like all outsourcing, a form of “social dumping” and unfair competition vis-à-vis businesses or corporations. We can see that workers in the network play the role of a reserve army of industrial labor, which weighs down wages, and that Amazon’s platform offers the same type of services as traditional service providers at an appreciably lower rate since it is not subject to the same regulations or to the same taxes.

With crowdsourcing, there remains a serious risk of turning human beings into a simple means to reach a commercial end, to turn them into a simple computer [SAG 11], to take away any sacred character, to see them as a simple raw material and end up in conflict with the moral philosophy of Kant who stated, “always treat others as an end and never only as a means”.

Crowdsourcing can be accused of being unfair. In one case, a team participating in the Shredder challenge organized in 2011 by DARPA (Pentagon), involving reconstructing documents that had gone through a paper shredder, was the victim of vandalism, since it was considered to be using unfair methods. The team used crowdsourcing in the form of puzzles while its competitors were using computer algorithms to assemble the images. The latter considered this method cheating compared to the algorithms that they were trying to develop, and quickly vandalized the crowdsourcing project.

As [FOR 11] emphasizes, the Amazon Mechanical Turk Marketplace is not a game or a social network but an unregulated market that pays no taxes and where workers, regarded as miroentrepreneurs, sell their labor for repetitive and unskilled tasks. They are underpaid[12], interchangeable, do not enjoy any protection and are doubly subordinate to the client and to the platform: in short, a kind of digital servitude. As [SAG 11] claims, it is probable that neither the "turkers" nor their employers declare their revenue, contribute to a social security or retirement fund and are listed in the business register. This off-the-books platform thus deprives States of lawful income and directly challenges their labor legislation. The fact of making anonymous people work without ever meeting them would encourage inhuman behaviors and exploitation of their workforce without limits or ethics. For their part, workers could also, for the same reasons as the employers, feel free from any moral obligations and develop cynical behavior [KIT 13] or fraud.

Regarding creative competitions that call upon "speculative work", i.e. work produced for free with the hope of being compensated [REN 14] by crowds of graphic designers who in the end have little chance of being to be paid, they greatly favor businesses that benefit from a much larger number of design proposals all the while having only a few individuals to compensate for a much lower overall cost than that of traditional agencies. It involves, finally, outsourced professionals rather than true amateurs. And, insofar as no contract connects the participant in the contest to the business, labor laws cannot apply; as much as it is a way of life for certain candidates, is it just a simple leisure activity for others. Crowdsourcing could also allow the renaissance of piecework and favors disengagement of the employer who would no longer be constrained to "be tied to a small number of people when one could have access to of a crowd of employees" [LEB 15].

With crowdsourcing, the consumption of free services on networks becomes a producer of data, information and value, making every aspect of social life productive, and free time and the consumption itself become production. In the same way, Guy Debord predicted "a colonization of every sphere of social existence by the authority of commodity in the organization

[12] The average hourly rate would be $2, according to [KIT 13].

of the Spectacle" [SAR 14]. In the continuation of the interest centered on the consumer through the economic model of "on demand", crowdsourcing appears to be participating in carrying out this colonization, and finally this integration of the consumer into the production process as an unpaid helper. As Harald Staun laments, downtime, free time, disappeared with the arrival of commerce and the profit motive during free time. Life itself thus becomes the engine of productivity, capitalism a "biopolitical" mode of production (according to Aspe in 2013, reported by [SAR 14]). Even our deepest human relationships are susceptible to being converted into algorithms by social networks and being valued commercially. Commercial relationships are also becoming widespread since, with collaborative consumption, each owner of a consumer item becomes a merchant who can rent out its use. The difference between production and consumption, between work and leisure is being blurred; Internet users create value through the free contributions that they provide and will be able to be reused and monetized via Big Data.

As certain authors [SCH 08] claim, Web 2.0 has all of the characteristics of an ideology, a totalitarian ideology promising "better tomorrows", an ideology that does not confine itself to the public and political sphere, respects no constitutional limit to its power, but interferes all the way into the private and intimate, the dream of a society where everyone would be connected, above nations and classes and within the framework of a worldwide government: in short, a Tower of Babel. Crowdsourcing could after all also show a kinship with libertarian and antiauthoritarian ideas since it substitutes activities of a community of volunteers that self-organizes in a decentralized way, for the hierarchical and centralized leadership of employees. Sociologist Michel Lallement who has studied Californian hackers thus believes that they are prolonging the libertarian counterculture [LAL 15]. The existence of the Internet seems to show the possibility of functioning that is harmonious and without hierarchy. In the participative encyclopedia Wikipedia, for example, an article written by scientist will find itself on the same level as an article written by a college student about his favorite comic-book hero. Linux is the result of the aggregated work of thousands of programmers, working on a common project on volunteer basis, for free and in a decentralized way. Christian Quest (*OpenStreetMap*) perfectly sums up this idea: "We would never ask you to have a master's degree in geography just to add a business near your house to an existing map!" Michel Bauwens uses the term "anticredentialism" for this type of position against the monopoly of degrees and laments the fact that we cannot be as credible as a scientist without a doctorate, or as a journalist without a press pass [BAU 15]. In the same way, some creative contests offer to graphic designers, artists, amateur publicists, beginners, without jobs and without references, to have as much chance of succeeding as an experienced professional with a job or a well-known personality and breaking into the industry more easily this way [REN 14].

The process of giving more responsibilities and power to act to the people and to the consumer, giving capacity for action to Internet users and freeing them goes back to the concept of *empowerment*. In this way, the participative science project fold.it states that its goal is for ordinary people to possibly, thanks to its puzzle game, be able to win the Nobel Prize [GOO 11].

Just as the border between production and consumption seems to disappear with crowdsourcing, the border between the authors who write and the readers who read is in the process of being gradually abolished since each one is now both reader and writer on the Web, therefore following even more Walter Benjamin's analysis (*Der Autor als Produzent*, 1934). Walter Benjamin thinks, in fact, that the emergence of new media would call into question the paradigm of the expert and that technological progress underlies political progress [DEO 14]. We could also talk about "active reading", a lack of separation between the actions of reading and writing, for example by annotating during the act of reading.

As we have seen in the preceding text, crowdsourcing is capable of appealing to Marxists as much as to liberals, for diametrically opposed reasons. As [SCH 05] remarks, for example, the ambiguity of info-communism is one of the principal resources of neo-liberal knowledge economy and can be described as simultaneously revolutionary and reactionary. It combines both the dreams of info-capitalism and those of Soviet constructivism. As Bastien Guerry also notes, "the 'leftists' of the Web are also liberals or even patriots" [BEN 14]. Elisabeth Grosdhomme Lulin also believes that "as far as ideas and doctrines are concerned, [the idea of the coproduction of a public service by its beneficiaries] has its roots as much on the left as on the right: on the left with self-managing utopias, on the right with libertarian utopias – on one side, in the wake of Pierre Joseph Proudhon, giving power back to the people, worker or citizen, to the other, following Friedrich Hayek, limiting the influence of the State on the economy and society". [GRO 13]. Rachel Botsman and Roo Rogers think, in the slide show *what is mine is yours*, that collaborative consumption responds both to socialist ideologies and capitalist ideologies without itself being an ideology. Finally, Nelson *et al.* [NEL 12] notes, for his part, that there is in the end, in all of this confusion, a paradoxical kinship between the Soviet ideology of Socialist Emulation and the liberal American ideas of gamification. Indeed, the Stakhanovist methods intended to motivate and develop worker productivity by rewarding the best workers with points, decorations, Soviet medals, the title of “Hero of Socialist Labor” and Stalin prizes, and by organizing competitions between workshops, businesses, factories, *kolkhoz*, *sovkhoz*, districts, towns, regions and republics to develop their spirit of initiative and enterprise, is in the end not that different from competing to be named employee of the month at the symbols of American capitalism that are McDonalds restaurants, which also sometimes offer gifts. The International Amateur Scanning League, a heritage digitization project using volunteers, offers, for example, medals based on this model, according to the number of DVDs burned.

This paradoxical proximity between socialist and liberal ideas is well represented in the “Californian Ideology”, which combines the hippie spirit of independence and autonomy and the Yuppie (Young Urban Professional) spirit of enterprise. Silicon Valley thus led to the emergence of libertarian ideology, liberal (*Yuppie*) and libertarian (*hippie*). Richard Barbrook [BAR 00c], the originator of the term, even believes that the Internet could be a modern form of the gift economy like sociologist Warren Hagstrom [BAR 00a, p. 504] who believed that science was also a gift economy [SUR 04]. Each contributor adds to collective knowledge and receives much more from other contributors than any one individual could provide. The scientific researcher, the developer or, more broadly, anyone who possesses information or knowledge does not lose anything by sharing it. According to [BAR 00b], Californian Ideology would nevertheless be the ideology of a kind of high-tech Nietzschean aristocracy, a kind of Jacobin elite, a cybercommunist avant-garde or of a kind of technocracy of the Web that he calls *digital literati* or *digerati*. These *digerati* are convinced that new technologies will revolutionize society. They seek to educate the masses and lead them toward modernity to create a utopian civilization, a society of information. The *digerati* would therefore be the reactionary modernists seeking to impose a renewed dictatorship of the proletariat which itself would last only the time necessary for the emergence of the new society. They are similar to Anonymous, whose slogan “we are legion” could also refer to the power of the crowds of Internet users who support crowdsourcing.

As [CAR 10] emphasizes, the Internet is the heir to the American libertarian and egalitarian counterculture and of the liberal meritocracy of the world of research and computing. They combine the ideas of Marshall McLuhan with certain radical libertarian ideas. McLuhan thought that the medium, that is to say the intermediary between the sender of information and the receiver that can take the form of speech print, film, radio, television and today the Internet, supersedes the content of the message itself ("the medium is the message" [MCL 68, p. 404]). McLuhan is also the origin of the concept of the "global village". Cyberlibertarians, these "technofans", proponents of the "myths of techno-utopia" [CHA 13], these partisans of technological determinism and "technological solutionism", believe that technologies will inherently bring about a democratic counterculture, change society and solve social or societal problems in the same way that Marxists were waiting for the Communist paradise of the development of productive forces and the revolution that they had to inevitably bring about.

1.5. Economic, sociological and legal consequences

1.5.1. *Economy of crowdsourcing*

On the strictly economic level, crowdsourcing could represent a significant source of markets and developments. Indeed, the cumulative time spent connected to the Internet worldwide should be close to 160,000,000 hours per day. The underlying idea of crowdsourcing is that the free time spent on the Web consuming content could be used in a way that is productive for the economy. In this way, personal data on the social Web are converted into statistical information, and therefore into value. Games on the Web in particular might have educational goals (serious games) but also produce data (gamification). Regarding crowdfunding in particular, according to the article "Global Crowdfunding Volumes Rise 81% in 2012" published on August 4, 2013 in *The Huffington Post*, crowdfunding sites had raised 0.89 billion dollars in 2010, 1.47 billion dollars in 2011 and 2.66 billion dollars in 2012.

1.5.1.1. *The disappearance of necessary work?*

Technologies evolve, productivity and growth increase, and the amount of work necessary for the survival of humanity has become lower and lower. In 1982, the United States was producing 75 million tons of steel with 300,000 workers. In 2002, 100 million tons were produced by only 74,000 workers. In the service industry, it is estimated that a traditional bank today requires ten times fewer employees to manage the accounts of the same number of clients. We are producing much more with much less work.

Jeremy Rifkin thus predicted that “it would require only 5% of the adult population to operate traditional industries” and that “factories, offices and farms, without workers or nearly so, will be the norm throughout the world” [RIF 96]. The improvement of labor productivity, linked to the development of new technologies, would destroy jobs. As an example, the traditional major industries are hiring a much larger number of employees than the largest Internet companies. For example, Facebook only had 3,976 employees in September 2012 for a billion users, which adds up to 250,000 customers per employee, Twitter had 900 employees for 500 million customers, Google 54,604 employees pour 1 billion unique visitors per month in July 2012 (according to Jean-Paul Lafrance, 2013, reported by [SAR 14]).

If technologies make it possible to reduce the marginal costs of services until they are practically free and if this movement has now also reached the production of goods equipped with sensors that produce data, these are the same bases of the capitalist economy, which will collapse according to [RIF 14]. According to these theories, if humans are replaced by robots or algorithms, they will no longer have the capacity, in the absence of income, to consume what machines have produced and we will be headed toward catastrophic overproduction crises calling into question the capitalist system and its Fordist model, which precisely hoped to avoid overproduction crises by indexing salaries to gains in productivity and in this way allowing workers to consume more that they have produced.

For Michael Osborne and Carl Benedikt Frey, cited by the blog InternetActu.net, 47% of the 702 professions studied could disappear via automatization[13]. This movement obviously would affect less skilled and less creative professions more, pushing employees toward higher activities. This movement would also be accompanied by the development of volunteer and associative activities, hobbies and crowdsourcing and comforts the proponents of an unconditional base income (or “universal income” or “guaranteed minimum income” or “universal allowance”) as we will see later.

At the beginning of the 19th Century, labor leader John Ludd destroyed numerous machines, and, throughout this period and up until the 20th Century, the working class contested the automatization of work and Fordism for the same reasons, without imagining that one day, the majority of workers

[13] See: http://www.internetactu.net/2014/06/17/travail-et-automatisation-la-fin-du-travail-ne-touche-pas-que-les-emplois-les-moins-qualifies (consulted June 23, 2016).

would be part of the tertiary sector of services. Today, the Internet and the Uberization of the economy can provoke the same type of anxieties. The famous economist Joseph Schumpeter's theory of creative destruction might, however, lead to optimism. This theory states that in the economy, the disappearance of industries goes hand in hand with the appearance of new activities participating in the evolution of the economy.

The current revolution in production process could nevertheless not be "Schumpeterian" and could destroy more jobs than it creates. As a result, according to Wendell Wallach of Yale University, 47% of jobs in the United States could be replaced by algorithms within 10–20 years. Moreover, the giants of the Internet are hiring only a few employees, when you compare their number to the company's total revenue and the number of its clients. On these subjects, we find an analysis that is relatively balanced between growth and decrease in postindustrial society from economist Daniel Cohen [COH 15].

1.5.1.2. *Crowdsourcing, basic income and the theory of the commons*

The invisible work of Internet users could be recognized in the form of a guaranteed income in order to restore to them the value that they have produced. Some proponents of a guaranteed minimum income, which would be financed by the value-added tax and paid out unconditionally and for life to citizens, even think that this "creative contribution" would make it possible to change their relationship with free time by encouraging the creation of businesses, but also unpaid labor, volunteering and allowing them to invest their time in contributory work in the service of others in the form of crowdsourcing, for example. For this reason, Bernard Stiegler prefers to talk about "contributory income" [STI 15]. Instead of working to earn an income and worrying about losing that work, a person would have an income to be able to freely devote themselves to the activity of his or her choice. No longer motivated by vital needs, but by higher needs, this would allow individuals to be more creative and innovative and to cooperate better. This movement would respond to the destruction of jobs by automatization and would not lead to a society of unemployed people, but a society of free and dependent entrepreneurs. This income would be the conceptual equivalent to that of copyright holders who make a profit from the commercial use of their ancestor's work until 70 years after his or her death. Citizens could consider that they profit from the accumulation of knowledge built up by humanity as intangible heritage. It could be seen as an investment by the government which allows contributors to pursue their participative work. With automatization, the evolution of work and the development of invisible volunteer work in the form of crowdsourcing, new forms of remuneration might emerge.

The concept of the commons comes from the 18th Century English countryside, which was seldom divided into separate properties and the use of which was shared among rural communities. Michel Bauwens, P2P theorist, remarks that "businesses base a portion of their economy on scarcity, which is contradictory to the logic of the commons" [BEN 14] and the work/capital contradiction is gradually being replaced by the commons/capital contradiction. These theories have, nevertheless, been criticized by the theory of the tragedy of the commons which says that free and open access to a resource fatally brings about its overexploitation and destruction. While its use is individual, its costs are collectively supported, and individual interest inevitably consumes beyond its needs. In the world of fishing, this phenomenon is demonstrated by fishermen who have an individual interest in taking as much as possible from the communal stock to the detriment of the collective interest and, in the long term, of their individual interest, once stocks of natural resources are exhausted. This inevitability requires the intervention of the State to prevent it from happening. However, in the case of digital heritage, the material is not "excludable": it is a non-rival good, the resource is not limited, its use by an individual does not prevent another individual from using it [PEU 12], sharing it does not consume it, does not threaten and does not divide up the resource, since, on the contrary, it can be multiplied indefinitely and for almost nonexistent marginal costs.

The emergence of virtual currencies such as BitCoin, created in 2009, could support crowdsourcing. These virtual currencies have, in fact, all of the characteristics of a crowd: decentralization and anonymity [LEB 15]. The Internet Archive, one of the major players in participative digitization, pays a part of its employees' salaries in the form of BitCoins. This virtual currency, convertible into dollars, makes it possible not only to transfer value from one Internet user to another and without intermediaries, but also to purchase Amazon gift certificates and consumer goods in certain marketplaces.

On May 1, 2013, already close to 300,000 BTC were in circulation at the unit price of 94.80 €. On March 13, 2013, the BTC was selling at $47 and in 2012 at only $4.93 (see: mtgox.com).

1.5.1.3. *The amateur, new motor of the economy and development?*

In the 17th and 18th centuries, the term *amateur* referred to those who could be elected to the Royal Academy of Painting without actually being painters, because of their passion for art. Today, it refers to people who act only out of love for a particular discipline, but it is also used pejoratively to discredit contributors for their lack of professionalism, also referred to as amateurism.

The figure of the amateur appears to be divided into two distinct types: one which will organize his or her professional and social life around a passion such that they do not impede it or even agree with it to the point of professionalizing it. This amateur regards the activity as sealed off from his or her professional and social life, even going so far as to sometimes do the activity in secret, hidden from familial or professional relations. There are both extroverted amateurs, who desire social recognition and follow a logic of networks, and introverted amateurs, who act more like selfless volunteers and respond to a sense of community.

With the development of crowdsourcing, we could move from a model of innovation, as described by Joseph Schumpeter, going from active producer toward the passive consumer, business being at the forefront of modernization and seeking to change users, to a model of innovation centered on active users who take their ideas back to the businesses that inspire them. The separation between production and consumption thus seems to be disappearing gradually. Users no longer want to be passive consumers and believe that a person does not really own something if they cannot open it up; they want to act and gather together in more and more in collaborative networks, they exchange, tinker and improve consumer goods through DIY (do it yourself), innovate and influence businesses without expecting anything from them in return other than the satisfaction of seeing their ideas come to fruition. They are all developing a "maker" culture. In the scientific field, we encounter, for example, a "garage biology", supported mainly by the association DIYbio (do-it-yourself biology). This type of association opens science and its means to amateurs. Innovation becomes the result of collaboration between producers and consumers who become its coproducers and coauthors.

As a result, according to Eric Von Hippel who talks about user-centered innovations, innovations through use or bottom-up innovations, 46% of American businesses in innovative sectors originated with a user: most of the time, a young person with a degree who benefits from a technical culture [VON 11] and who cannot find the service or the product which he or she needs on the market, since traditional companies are still not organized for custom manufacturing or ready to risk investment in the face of uncertain demand. This enthusiast is usually ready to invest time and money to develop, to build rather than to buy, to produce rather than consume and is ready to share his or her discoveries for free. He or she has access to more and more advanced computer science and technologies and the production of a prototype is less and less expensive. The skateboard was invented by consumers in this way, and 80% of innovations in scientific instruments were developed by users and we no longer count developments, which are the fruit of users in the world of freeware [VON 05]. In the area of libraries, the LMS and OPAC functions produced by service providers have therefore largely been inspired by clubs of users made up of librarians, as Von Hippel always points out. Generally, lead users tinker with a product for their needs, this product is adopted, copied and improved by other consumers and the success is such that businesses have ended up becoming interested. Well beyond traditional market research, businesses would therefore have every interest in anticipating and collaborating with these lead users by offering them toolboxes, forums, social networks and platforms. According to [VON 11], consumer–innovators still only represent 6.1% of the population aged over 18 in the United Kingdom, 5.2% in the United States and 3.7% in Japan.

With the development of freeware in particular, users are being recognized more and more as a possible source of innovations. The company Dell, for example, has launched the side site Ideastorm and has collected more than 10,000 proposals for ideas for improving its products and services. With its Techshop project, fully open to the ideas of consumers, the company Ford has increased its filing of patents by 30%.

By comparing the ideas that come from professionals with those coming from users, Poetz and Schreier [POE 12] unsurprisingly report that, according to his study, the ideas from users would be more innovative (average grade of 2.6 versus 2.12) and more advantageous for consumers (average grade of 1.86 versus 2.44), but also somewhat low in terms of feasibility, as the ideas of professionals have a tendency to be much easier to carry out (average grade of 4.33 versus 3.91).

The history of science is moreover full of inventions coming from people outside of the field who are not seeking to reproduce the established models with which they were trained and who are likely to cause innovative ruptures. In the new economy, it in fact seems that businesses need to increasingly connect to external ideas and energies and integrate the consumer into the production process [LIG 12].

Crowdsourcing makes it possible to create an ecosystem of innovation by having people with very different skills and backgrounds work on common projects with the help of new technologies.

1.5.2. *The users of crowdsourcing*

Crowdsourcers, clickworkers and other “web proletarians” are sometimes compared to “Oompa-loompas”, a tribe that works by having fun for the chocolate maker Willy Wonka in exchange for chocolate in Roald Dahl’s novel, *Charlie and the Chocolate Factory* [REN 14]. Do all these workers constitute a socioprofessional category or even a social class? Will we see the emergence of a class of *prosumers* or *produsers*, that is to say, a class of individuals who are both producers and consumers–users of their own products, more tied to the shared use of goods than to their private appropriation, as Jeremy Rifkin predicts?

The emergence of Generation Y or *digital natives* in business could also have an influence on the development of crowdsourcing. This generation is overturning hierarchies, authorities and frames of reference; its culture is more open and participative. It is therefore probable that it will be more open to crowdsourcing, as shown in figure 1.10, which shows the percentage of contributors to Wikipedia by birth year.

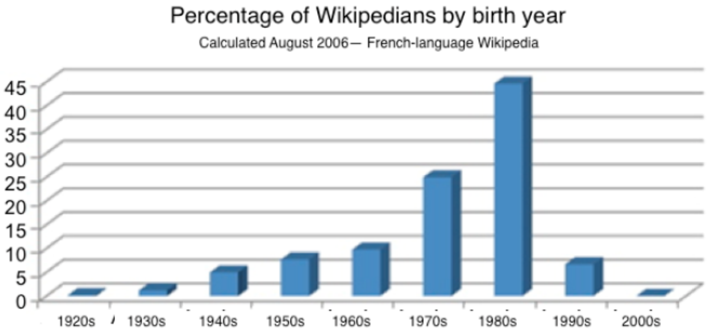


Figure 1.10. *Percentage of Wikipedians by birthdate, according to Wikipedia*

Younger generations have a tendency to build their identity through their participation on the Web, for example, writing blogs, setting up or participating in fan forums. Amateurism and crowdsourcing could be a way for them to cultivate their e-reputations and showcase their participation in cultural projects: participation which could also be beneficial for their resumes and job search.

However, older generations can also be involved in crowdsourcing. Although the Beuth Hochschule Für Technik's [BEU 14] report shows that 26% of Wikipedians are between 22 and 26 years old, it also indicates that 28% of them are over 40 years of age and that 36% of this older category is some of the most active. In fact, many are retirees who have significant free time available and are already somewhat active in organizations and volunteering and, when it comes to crowdfunding, they have more available capital than the generations preceding them, who often need to pay off loans and less often have access to real estate income and financial investments.

The free time that all of these categories of populations have is a formidable reservoir of goodwill for crowdsourcing. Each minute, 35 hours of video are put on YouTube in this way, and each hour 38,400 photos are posted on Flickr, according to the report "Crowdsourcing in the cultural heritage domain: opportunities and challenges"; [PAR 13] mentions the figure of 72 hours of video, uploaded to YouTube each minute and 2,500 photos on Flickr.

The theory of communities is a conceptual framework that can be enlisted in order to analyze crowdsourcing from a sociological perspective. An online community or a virtual community of practice is a relatively homogeneous group of Internet users who work together for the benefit of a common undertaking in a relatively self-organized and informal way, who help each other in order to resolve practical problems in the form of mutual commitment and who share a repertory, i.e. a heritage of information [WEN 98]. The virtual community is generally built around a solid core of active members. A culture of community is likely to develop with a common identity, shared references and implicit rules. This culture is transmitted to novices by leaders or by senior members of the group [DAE 09]. When this community is more heterogeneous, is more creative, produces new knowledge recognized, is considered authoritative in the scientific community and wants to ensure that this knowledge consists not just of improving practice, but that it is "usable knowledge" [MEY 11], that is to say, having an influence on public policy, we instead talk about epistemic communities [MIL 11]. A community of practice can gradually transform into an epistemic community [LIE 14a]. The type of community will obviously be variable depending on the type of crowdsourcing present. Epistemic communities are found, nevertheless, more within the scope of citizen science or Wikipedia than in the digital libraries projects that we have identified.

1.6. Managerial, library science and technological consequences

1.6.1. *The cultural factor*

Very broadly speaking, in Anglo-Saxon culture, the sharing of information by communities of interest is relatively natural. This is still not always the case in Latin or African institutions, for example. The latter would have to go through a major cultural change in order to adapt to these new models.

Meanwhile, it seems clear that the cultural factor is important in the adoption of crowdsourcing [EST 15]. Moirez [MOI 13c] notes, for example, that Web 2.0 projects and in particular, crowdsourcing projects in libraries, for example, are more successful in Anglo-Saxon countries due to cultural differences. Bœuf *et al.* [BOE 12] notice the same difficulty in developing citizen science in Latin countries due to lower involvement of individuals from Latin culture compared to individuals from Anglo-Saxon cultures in collective life and due to a larger distrust, a fear of being taken advantage of or of working on a useless project.

1.6.2. *The corporatist factor*

Libraries have gradually seen storekeepers challenged with the development of open access, catalogers called into question with the development of shared cataloguing on a worldwide scale, and finally, reassessment of acquisitions with the development of electronic periodicals followed by e-books. As Clémence Just reported in *Archimag* on July 21, 2015, researchers at Oxford University estimated the probability that the profession of librarian would be automated soon at 64.9%.

Libraries sometimes remain distant from the world of business and their managers often consider public interest to be more ethical than the profit motive. Crowdsourcing could therefore be considered a form of privatization or as a renewed and alternative public/private partnership [MCS 11].

It could be difficult for the profession of librarian to experience challenges to its monopoly. Libraries are no longer the inescapable intermediary between information and the public. This feeling, which is never clearly expressed, is similar to the feelings of gatekeepers, the guardians of the established cultural and political order, described by Cardon [CAR 10]. According to this researcher, the media seeks to retain its privileges, control and monopoly on access to information by a people considered insufficiently responsible and enlightened to form an opinion independently. The Internet is therefore, for them, a threat to the vertical and monopolistic model of diffusion of information that they have created and a challenge to their authority. The information which was produced by some (including authors, journalists, editors and librarians) in the Web 1.0 is now produced by the multitude with the Web 2.0.

For a cultural institution such as a library, agreeing to open up its indexing, its cataloging, its choice of which documents to digitize to amateurs requires a major cultural evolution[14]. It involves, in fact to going from a policy of supply centered on collections and the activities of librarians to a policy of demand centered on services, the needs and activities of users then directly activated and driven by the initiative of the individual user themselves and which corresponds well to "on demand" models. The user thus becomes a central actor in the digitization policies of libraries, hitherto reserved for its professionals [KLO 14]. According to this point of view, as depositories of printed heritage, libraries should become actors in the development of heritage that includes Internet users.

For professionals, the setting up of a crowdsourcing approach in a cultural institution can nevertheless, justifiably, be felt as devaluing the work of curators and documentalists, which could lose value since they can be done for free and by anyone. This change therefore requires a significant investment in change management and in internal communication. As Ben Brumfield reports on his blog, manuscripttranscription.blogspot.fr, as part of the Manuscript Fragments Project developed by Harry Ransom Center, close to 20% of comments (around transcription of sources or the identification of fragments) received about medieval manuscripts came from professionals, but these all preferred to send e-mails rather than contribute directly online.

[14] It nevertheless remains important that institutions be guardians of permanent references and it is certainly possible to create hybrid information systems that make it possible to add to permanent references without modifying them.

This can probably be explained by the necessity of preserving their reputation and the fear of putting their contribution at the same level as that of the layman and to see their skills discussed by them or their authority shared with them. Crowdsourcing is generally seen by professionals as a simultaneous loss of control over the choice of documents which will be digitized (in particular with digitization on demand) and the way in which the cultural material will be exhibited and used and, at the same time as an inescapable commitment to contributors, the results of their work having to be accessible permanently.

Nevertheless, the involvement of private amateurs can, in certain subjects, provide contributions for the benefit of public institutions and usefully complete the work of professionals whose workforces, means and knowledge remain limited despite all good intentions. Thanks to crowdsourcing, libraries can draw from an unlimited crowd of Internet users which can contain real specialists in a particular subject who know the content and the interest of a particular book much better.

Crowdsourcing is a more user-centric model. Successor to the Web 2.0, it is more interactive and reciprocal and less hierarchical than top-down models of diffusion of knowledge. Nevertheless, certain institutions are still sometimes not sufficiently centered on their users and remain focused on supply. They can sometimes not worry enough about demand. Yet, as [LEV 14] mentions, cultural institutions could from now on concentrate on the aggregation and delivery of digitized heritage while Internet users could take care of the enrichment of metadata. This would involve a cultural revolution in the profession, since archivists privilege collections and their meticulous description compared to their users. It would involve, on the contrary, favoring user access to content, even if it has not been described yet, and it would be free of charge. Nguyen *et al.* [NGU 12] thus invite librarians to give more power to their readers to encourage their participation and to develop a true culture of participation.

With crowdsourcing, if the Internet user becomes a librarian, the librarian and the library curator could feel brought down to the level of Internet users. Yet, certain companies are based on the aristocratic idea of election or delegation. Each domain has its specialists, its experts whose legitimacy and authority might now be called into question. The profession of curators and librarians could not be an exception and crowdsourcing could be the name given to their challenge by the mass of anonymous and sometimes incompetent Internet users: the name given to their “Uberization”.

1.6.3. *The reign of the amateur: toward mediocracy?*

Loss of monopoly, of control, of power; the risk of low quality, malicious intent, vulnerability to *lobbies* and ideologies; risk of non-representative minorities taking control, questioning professional expertise; loss of responsibilities, etc. There is no lack of reasons to oppose the coming rapid expansion of crowdsourcing in libraries.

Indeed, professionals and experts who have produced metadata within a formal framework that is institutionalized and collective and recognizes risks may not favor diffusion on the Web leading to participative redocumentarization. This can be synonymous with personal and individual appropriation of collective heritage by a handful of Internet users who feel authorized to leave their traces, to tag, or to add their profane, informal, personal, intimate, banal, average, trivial and mediocre points of view.

The comments on the images are often ones like “Excellent”, “Superb”, “WOW!” “Great!”, “Perfect!”, etc. [LIE 14b] and are, consequently, completely unusable.

The term “amateur” itself is ambivalent. It can refer to both someone who loves something or a non-professional who does a bad job. The amateur is an enthusiast who dedicates a large part of his or her time to the passion and who does not look for any compensation other than recognition. Fundamentally, what distinguishes the professional from the amateur is the knowledge of the methodologies and standards for cataloguing and bibliographic descriptions, the rules for indexing or diplomatic transcription, TEI or EAD encoding standards, etc. Allowing access to this knowledge by the layman and neophyte could end up devaluing these skills and expose the fact that this knowledge is not based on any specific science, but on a group of rules that can also be partially arbitrary. By accepting that amateurs are capable of acquiring their knowledge, professionals might however convert amateurs into semiprofessionals and therefore into defenders of their professional interests.

As Rose Holey emphasizes, when libraries were still only offering their users printed documents, readers already enjoyed interacting with the reference librarian or with other readers and shared documents, but it was not possible for them to annotate a book under penalty of exclusion. When libraries went from printed books to electronic libraries, readers need to be able to not only be only simple consumers of information, but also be producers and above all collaborators for information professionals. Thus, they can now add a summary of the book or article that they have read; share it with their social networks; add information, metadata, comments, annotations; correct errors in metadata; converse with other users and even organize collaborative work with them. Amateurs and professionals can now collaborate, as much as amateurs who collaborate can acquire a high level of expertise rather quickly. By agreeing to use these crowds of amateurs, libraries might then be able to find more easily an expert on a particular subject within reduced teams of curators [HUV 08]. In any case, crowdsourcing already has a huge advantage compared to the traditional outsourcing already widely practiced by libraries, which have access to low-cost labor such as from India, Vietnam or Madagascar, since a knowledgeable and passionate genealogist generally becomes more competent more quickly and knows the subject better than a subcontractor from a low-cost country whose language and culture are more distant and who might only work on the project for a short period of time.

1.6.4. *Crowdsourcing: the highest stage of outsourcing?*

As we have previously and extensively mentioned, crowdsourcing falls within the economic movement of flexibilization and outsourcing that began with the subcontracting of entire facets of production to countries with more competitive labor costs or to suppliers, consultants, or even sometimes employees of the business who have become self-employed workers or miroentrepreneurs. With crowdsourcing, we now outsource on the Web. Some people even talk about “open outsourcing”. Instead of outsourcing to a specific subcontractor in a country for a low cost, crowdsourcing is outsourcing to a crowd of anonymous Internet users from every country.

In a difficult climate for libraries, crowdsourcing can also prove to be a way to do more with fewer resources. In the area of digitization, in particular, we have witnessed in recent years outsourcing of OCR correction or metadata entry work to low-cost countries (Madagascar, India, Vietnam, etc.). This outsourcing has made it possible for digitization service providers to reduce costs and offer more high-performance services by developing abroad, using foreign companies that already exist or benefitting from specialists that certain foreign countries have access to. Outsourcing is also an occasion for a company to open up to other work cultures and to enhance its own procedures. Crowdsourcing is a form of outsourcing that is not concerned with where the contributor works, as the only condition required is to be connected to the network of the WorldWideWeb. Indeed, Jeff Howe in the article "The rise of crowdsourcing" published in *Wired Magazine* in 2006 and which popularized the term crowdsourcing, pointed out that during the ten previous years businesses have sought to relocate to countries where labor was cheaper such as India or China, but that the place where the employees are located could have less and less importance in the future, insofar as they are connected to the network. Indeed, why relocate to low-cost countries when, via networks, it is now possible to mobilize, for very low or no cost, a more diverse, motivated, qualified and competent labor force? For libraries for example, this diversity is a major asset, since it has become possible to benefit from the skills of specialists in a particular domain well beyond the narrow limits of teams of curators who, despite a good general education, can never be specialists in every discipline. It allows, moreover, the development of multidisciplinarity. As Nicolas Colin claims in a statement reported by the blog *Internet Actu*, "there is now more power outside of organizations than inside them". Crowdsourcing also thus poses the question of the borders of the organization since it makes it possible to create value beyond those borders [LEB 15, REN 14b].

The border between what can be done by the artificial intelligence of machines and that which needs to be done with human intelligence is perpetually changing. For the moment, the work entrusted to human beings is only done so because it cannot be entrusted to machines.

However, this outsourcing could also be a first phase of the suppression of certain cultural public services after having demonstrated its own feasibility and after having reduced them to a form of begging on the Web. Indeed, in a context of disengagement by the State, why continue to pay professionals to do work that amateurs are willing to do voluntarily? Crowdsourcing could therefore amount to a form of Uberization of public services. With crowdsourcing in libraries, we might be witnessing an "Uberization" of libraries, that is to say a replacement of the services provided by a professional with that of an amateur. Like other forms of Uberization, it could therefore also provoke hostile reactions.

In this conceptual chapter, we have defined crowdsourcing in particular in its application to digital libraries. We have also detailed the historical and ideological origins of the model, and its economic, sociological, legal and managerial consequences.

In the chapter that follows, we will illustrate crowdsourcing in digital libraries with a panorama of the most representative projects in each major type of project according to the type of work that is required of Internet users: uploading, digitization and print on demand, OCR correction and indexing.

2

Overview of Several Crowdsourcing Projects Applied to the Digitization of Libraries

In this overview, we provide synthetic analyses for the major types of tasks that can be entrusted to Internet users. In order to demonstrate these types, we have also selected a single project representative of each type. We thus distinguish uploading and participative curation, digitization on demand through crowdfunding, printing on demand, participative optical character recognition (OCR) correction and folksonomy.

2.1. Putting content online and participative curation: the Oxford's Great War Archive and Europeana 1914–1918

Uploading content online and participative curation consist of allowing Internet users to complete institutional digital collections with their own copies or selections.

In 2008, Oxford University in the United Kingdom created an archive containing 6,500 digitized images thanks to the contribution of English citizens who provided their personal archives of the Great War, their family letters, photographs and war souvenirs for digitization, with notices written by the general public. These documents from private archives made it possible to create a public collection.

The success of this project encouraged Europeana to mobilize other national and local institutions throughout Europe in a partnership with

Digital Libraries and Crowdsourcing, First Edition. Mathieu Andro.
© ISTE Ltd 2018. Published by ISTE Ltd and John Wiley & Sons, Inc.

Oxford University. As a result, “Adding your story to Europeana 1914–1918” was inspired by this initiative to collect memories of the Great War in several European countries.

In France, from November 9 to 16, 2014, more than 70 collection points throughout the country conducted a similar operation with the help of volunteering institutions, which were able to participate in this “big collection” by making personnel and digitization workshops available.

Here, we will mention only the co-construction of digital libraries and not the co-construction of physical collections and participative acquisition of printed books by committees of users. Our scope includes only digital libraries.

The possibility of Internet users and volunteers completing public collections with the digitization of their own heritage collections calls into question the concept of collections resulting from selection by librarians. Opening up libraries’ collections management policies to Internet users thus represents a major evolution in their mission. Placing content online and participative curation are similar to digitization on demand through crowdfunding, which will be addressed in Chapter 3, in the sense that the Internet user becomes an actor in the collections management policy and the building of collections, but, unlike crowdfunding, this participation stops at document selection or making documents available and does not go as far as financing of the digitization itself.

For a more exhaustive overview, we could also mention Internet Archive, Wikimedia Commons, Picture Australia, Wir waren so frei, Open Call – Brooklyn Museum, Make history, Click! A Crowd-Curated Exhibition, The Changing Faces of Brooklyn, ExtravaSCANza, etc.

2.2. Digitization on demand in the form of crowdfunding applied to digital libraries: the European eBooks on Demand network

Crowdfunding is generally considered a form of crowdsourcing that relies not on the work and intelligence of Internet users, but on the financial resources of these crowds [ONN 14]. Brabham considers crowdfunding, on the contrary, more as a means of alternative financing which, unlike crowdsourcing, does not allow Internet users to weigh in on the politics of the project. Nevertheless, as [ONN 14] shows, crowdfunding projects often finally also call on forms of crowdsourcing (vote, help with promotion on social networks, etc.) so that the project is "carried away by the crowd".

According to [ONN 13], who provides a definition of it, crowdfunding "consists of a project leader (whatever their status: private citizen, commercial or non-profit organization, etc.) calling upon the services of a financing platform (generalist or specialized in order to propose a project (completed or not) to a community (large or targeted) of contributors referred to as *backers* in exchange for possible compensation defined beforehand". This community is usually recruited via social networks. We can cite, for example, the platforms Ulule and KissKissBankBank that occupy a sizeable position in this market.

From the point of view of libraries, digitization on demand is above all a service provided to the user, but it is also, for libraries, a means of outsourcing the financing of their digitization and a way to complete their digitization programs. The financing for digitizing a document can be individual or collective. It can be motivated by the individual need to gain access to a document that is difficult to access, by the need to financially support an institution or by the desire to increase accessibility, and therefore promote a particular work.

It is therefore clearly a form of crowdfunding applied to the digitization projects of libraries.

The European *eBooks on Demand* (EOD) network, launched in 2006 as part of the European eTEN (2005–2008) project, and led by the library at the University of Tyrol, allows the libraries participating in it to use a payment platform to set up their digitization on demand services.

Each library is invited to add EOD buttons dynamically to the titles eligible for digitization in its online catalog, that is to say, before a given date, usually 1900. Users interested in a particular title in the library catalogue can push these buttons and be sent to a payment platform in order to obtain the PDF of the corresponding paper copy. The digital documents are then made available, after an average time of a week, via the Digital Object Generator that generates a multilayer PDF with OCR and a cover presenting the service. After a period of around 2 months of embargo during which the sole beneficiary of the digital document (6 months at the National and University

Library of Slovenia), the library makes the digitized document available in its digital library and adds a link to the record to its online catalog.

The service is carried out by the digitization services of libraries, not by an external service provider, although this is not necessarily always the case[1]. This has the advantage of allowing for lower prices. The working time of public servants is not usually completely reflected in the price, but this also has the inconvenience of forcing libraries, which would like to participate in the project to invest in digitization workshops that require expensive equipment and a qualified staff. However, it seems that there are now examples of libraries using a private provider.

In 2009, there were 13 institutions participating in the network. Today, 30 libraries in 12 European countries participate:

– Austria: University of Innsbruck, Library (coordinator), University Libraries of Graz and Vienna, Library of the Medical University of Vienna, Vienna City Library, Saint Pölten Diocese Archive;

– Germany: University Libraries of Regensburg, Greifswald, Leipzig and Humboldt-Universität zu Berlin, Bavarian State Library, Saxon State and University Library of Dresden;

– Denmark: The Royal Library;

– Estonia: National Library, University Library of Tartu;

– France: Bibliothèque interuniversitaire de santé (BIUS), Bibliothèque nationale universitaire de Strasbourg (BNUS);

– Hungary: National Széchényi Library of Hungary, Library of the Hungarian Academy of Sciences;

– Portugal: National Library;

– Czech Republic: Moravian Library in Brno, Research Library in Olomouc, Library of the Academy of Sciences in Prague, National Technical Library;

– Slovakia: University Library in Bratislava, Slovak Academy of Sciences;

[1] With the notable exception of the Bavarian library which already uses an external service provider.

– Slovenia: National and University Library;

– Sweden: Umeå University Library;

– Switzerland: The Swiss National Library.

In total, 3.5 million records of works have been offered. The members of the network pay an annual subscription of around €1,000 to the library in charge of coordination. These fees cover the actual costs of administration, OCR, access to the Order Data Management payment platform, its maintenance and its help service. Since these infrastructures are mutualized, their costs are thus shared.

Between 2007 and 2009, according to [GST 09], 3,200 books (840,000 pages) were digitized by 2,000 users. Between 2007 and 2011, according to [GST 11], close to 5,000 books were digitized, with around 1 million pages scanned. Close to 2,500 people placed an order over this period. If we consider the three libraries that have enjoyed the best results, each one of them receives an average of between 250 and 350 digitizations of books per year, which is one per day from the opening of the library.

Taken as a whole, the number of orders and the revenue generated by EOD are growing, as this extract from an activity report demonstrates.

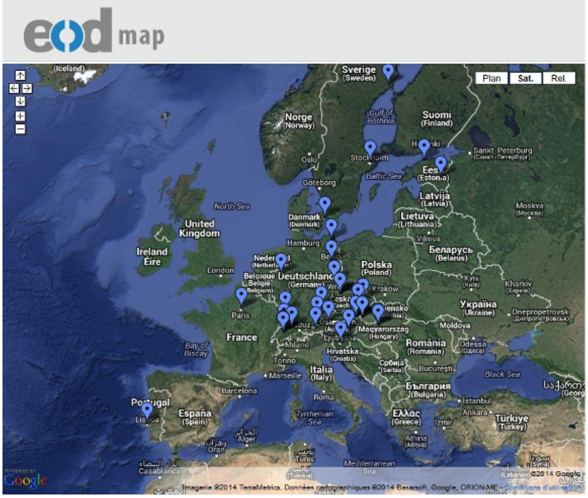
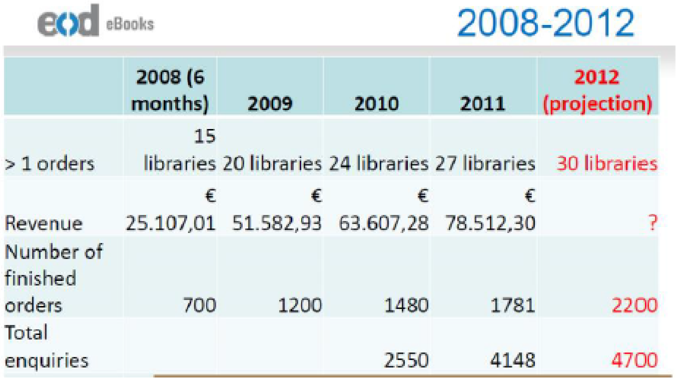


Figure 2.1. *Location of the members of eBooks on Demand network on July 8, 2014, from https://www.facebook.com/eod.ebooks/app_402463363098062 (consulted June 23, 2016). For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*



eBooks 2008-2012

	2008 (6 months)	2009	2010	2011	2012 (projection)
> 1 orders	15 libraries	20 libraries	24 libraries	27 libraries	30 libraries
Revenue	€ 25.107,01	€ 51.582,93	€ 63.607,28	€ 78.512,30	?
Number of finished orders	700	1200	1480	1781	2200
Total enquiries			2550	4148	4700

Figure 2.2. *Extract from an EOD activity report, from [KLO 14]*

In the majority of cases, the flat rate was €10 for the user for management costs. Added to the cost of these 10 euros is a per page cost that varies between €0.15 and €0.30. In the end, the price of a book digitization usually falls within a bracket of €20–€49. A minority of digitized books (20%) cost more than this bracket. A digitized document that is 250 pages long would thus cost the Internet user between €30 and €130 [MUH 09]. And, in 2009, the average cost for the reader would more precisely be €53.

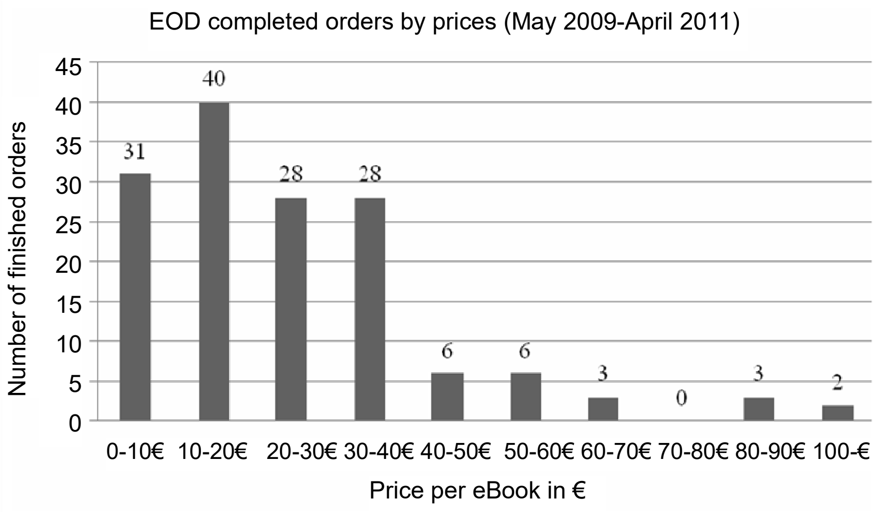


Figure 2.3. *Orders per price class during the 2009–2011 period at the National Library of Slovenia, from [BRU 12]*

According to a survey conducted by EOD in 2009, these rates are considered high or very high by 30% of respondents, but 95% remained satisfied by the quality/price ratio. However, beyond €50, Internet users would be more reluctant to finance the digitization of a book [MUH 09].

At the National and University Library of Slovenia, for example, between May 2009 and 2011, orders seem to mostly involve relatively low-priced classes.

According to our calculations[2], the average price for digitizing a 200-page book (with OCR) would therefore be €46.24, by calculating the average rate of the rates of the 36 partners. The survey previously mentioned also shows us that the multilayer PDF with OCR is, along with possession of the original book, the preferred format for users, ahead of the PDF image, the online OCRized text and reading the printed original in a library.

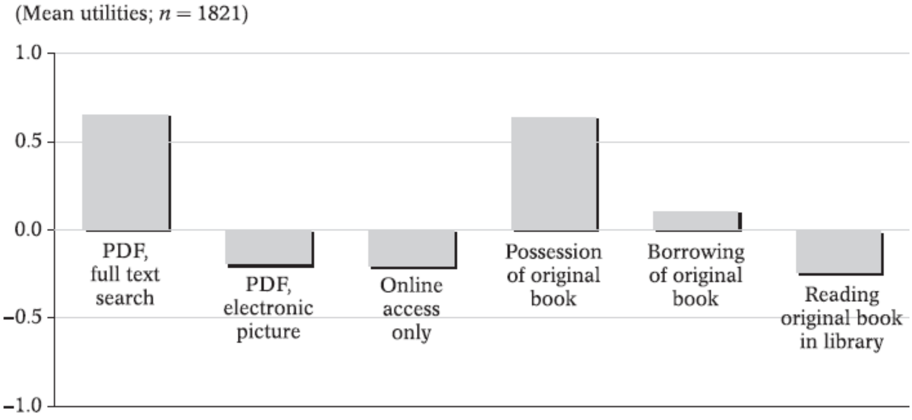


Figure 2.4. *The form in which users prefer to consult documents, according to the survey related by [MUH 09]*

According to the survey related by [MUH 09], a delivery time of more than 3 weeks would be viewed very negatively by users, as shown in Figure 2.5.

[2] According to the EOD rates available at: http://books2ebooks.eu/fr/prices (consulted on July 9, 2014).

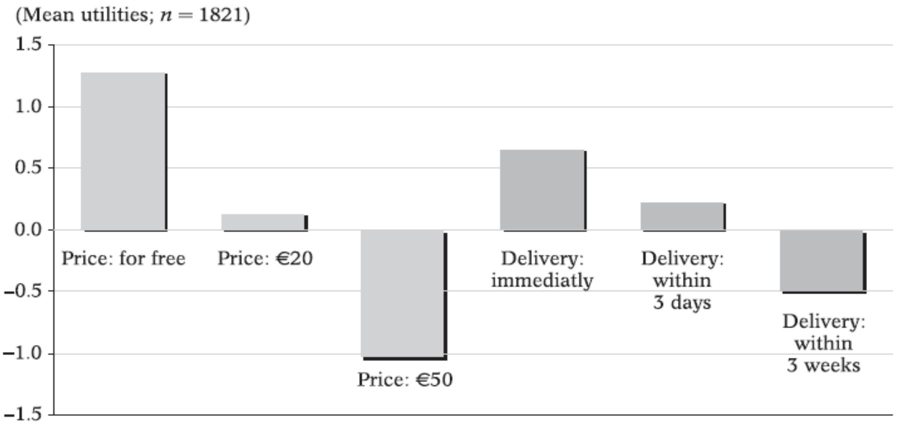


Figure 2.5. *Positive/negative perception according to prices and delivery times, according to the survey related by [MUH 09]*

Users are mostly motivated by reasons related to work (over 60%), but 16% of them are bibliophiles, amateurs or collectors.

The demands were mostly justified by the fact that users have no other way to procure documents and by the difficulty of accessing some old books.

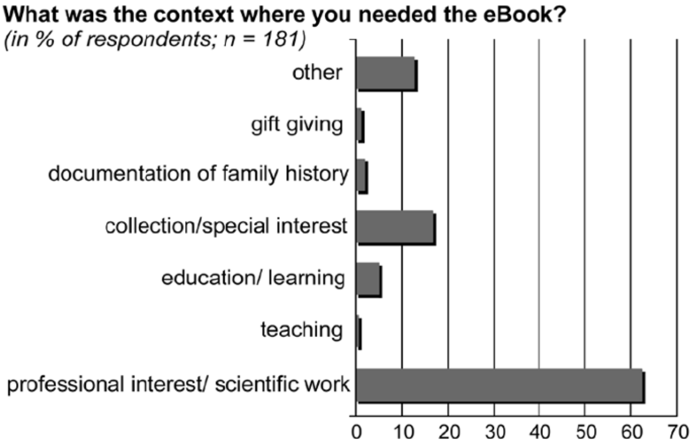


Figure 2.6. *Areas of interest for users, from [GST 11]*

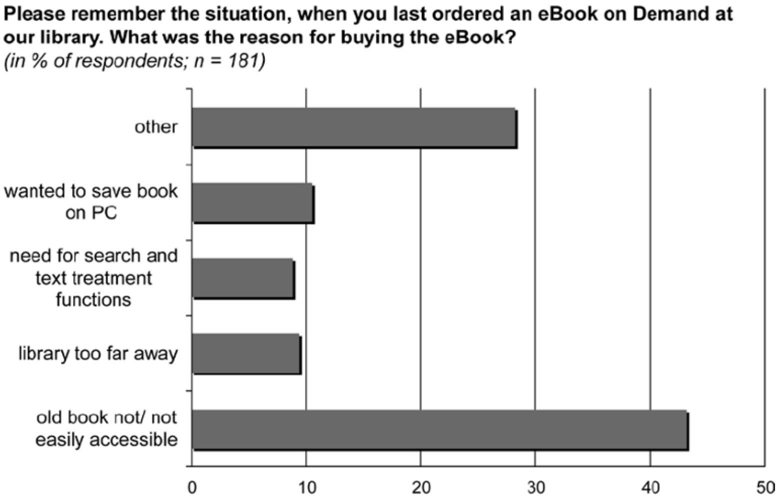


Figure 2.7. *Reasons why users placed orders, from [GST 11]*

It was mostly men who placed the orders. The clients were usually from the same country as the library from which they are ordering the digitization. Thus, according to statistics from the Bibliothèque interuniversitaire de santé, France is the leader in terms of demand.

Geographical area	Number of users	Number of orders
Europe	92	299
Same country as the library	155	568
Worldwide	51	160
Total	298	1,027

Table 2.1. *Statistics of EOD orders from the Bibliothèque interuniversitaire de Santé, from [KLO 14], translated by us*

From 2009, if the institution chose this option, the digital books were also sent to Amazon Booksurge accompanied by metadata and ISBN. *Print on Demand* (POD) reproductions can thus be ordered via Amazon Booksurge and POD books are also available directly for purchase on Amazon.

Digitization on demand of documents can be offered to Internet users from several sources:

– certain Internet users, while searching for a document, will enter its title directly into a search engine and arrive at a well-referenced platform on which the financing of this digitization will be offered to them;

– others will receive this offer from buttons on the bibliographic records of online library catalogues. These buttons could then offer to “download”, “digitize”, “reprint”, and/or “send” the document depending on the services available [STA 13];

– some contributors could be directly solicited from digital libraries. These could then display not only the documents that they have digitized, but also those they would like to digitize and offer financing to Internet users. This function would also allow libraries to exhibit their future digitization projects.

This type of service would allow libraries to offer their readers professional quality reproduction and digital reproduction services, very often nonexistent, and to thus respond to their users’ demands for reproduction without having to bear the cost. In addition, it would make it possible to modernize interlibrary loan services, which have become very out-of-date, and to make them more effective. It would also allow libraries to complete their digitization programs all the while sharing their policy for selecting documents to digitize with the general public and various patrons.

In fact, this economic model could also interest institutions, foundations and patrons, in addition to the Internet users who would be provided with an access service to documents. The acknowledgment “this book has been digitized thanks to the support of Mrs. X, Institution Y, or Foundation Z”, following the Google Adwords economic model, would foster encouragement for this type of participative financing. A book which generates, for example, 6,000 visits would provide web traffic whose value could be estimated at the cost of a digitization and it could therefore be profitable for an investor to finance its digitization, especially since, unlike in a Google Adwords type advertisement, its duration would not be limited and the price to be paid would be fixed. Once the digitization has been funded, the name of the investor will remain displayed with no extra cost no matter the number of visitors or number of clicks generated. For certain documents, it is, in fact, probable that beyond a certain number of visitors, the digitization costs would be recouped by a return on investment, via ads seen and the Web traffic generated by links. Digitization of cultural heritage could thus, from a certain point of view, also be considered an investment, a communication and marketing strategy. Indeed, a digital library makes it possible to reach Internet users who are interested specifically in a particular subject anywhere in the world. It would be enough for businesses wanting to reach a specific category of Internet user to finance the digitization of books that interest them. By providing this service, they would improve their image and could also gain web traffic via links pointing to their sites.

By opening up the collections management policies and acquisition policies of their digital libraries, libraries can develop digital collections directly chosen by their users. While satisfying a documentary need, the Internet user also sustains the digital library, which becomes a cocreation. The digital library's collection management policy and acquisition policy are thus determined by its users who decide which titles will go into in the digital collection, and not by professionals who may not have mastered every area of knowledge. Who better than the users themselves can know the users' needs? The digital library built this way over the years by digitization on demand is therefore the reflection of the choices of the Internet users who preside over its acquisition policy. It is a digital library enriched by Internet users for Internet users. It is the work of Internet users themselves, in a *bottom-up* model that is radically different from the logic of supply disconnected from demand, the public and use. Behind the concept of digitization on demand, like printing on demand, lies the concept of an immediate, real-time start of the supply by the individual user to a service provider who maintains a potentially accessible service. Digitization on demand is centered on the user and on his or her personal demand. Its model is not *top-down*, but *bottom-up*. It responds particularly well to the needs of students and teachers who increasingly work remotely and may need to access documents from everywhere and at any time [TAF 11]. Digitization on demand is also the evolution of a *just in case* economic model (stocking to anticipate demand) which was that of libraries (building collections to anticipate the needs of readers) toward the *just in time* economic model [REI 08]. Yet, libraries draw a large part of their expertise from building collections and in their acquisition policies, and might therefore feel themselves directly challenged by this economic model which can only become established with change management. This movement continues the movement that led libraries to adopt the "self-service" model, originally developed in stores, in the form of open-access collections". Indeed, it already involves subcontracting the shopkeeper's work to the reader him or herself and, in a certain way, to gradually integrate the consumer into the production process.

By subscribing to this type of approach in the form of digitization on demand, libraries outsource to the general public the laborious work of selecting documents, which still deserve to be digitized and the thankless work of verifying the documents that have not yet been digitized, since it is unlikely that an Internet user is financially able to fund the digitization of a book that has already been digitized for a cost which remains considerable. This identification work can be difficult to automate and, as Pignal and Perez [PIG 13] point out, "in absorbed cost, selection can be more burdensome than the digitization of an entire collection".

Because of digitization on demand through crowdfunding, public funds could be concentrated on the documents that are not likely to interest private citizens and which have a heritage, scientific or historical interest. Public funds could then be used better and private funds could take over the digitization of books that interest private citizens or, if they are likely to generate web traffic, interest investors or patrons. Digitization on demand could provide an alternative model to that of mass digitization offered by Google Books [CHA 12]. Mass digitization with public funds and individual digitization on demand with private crowdfunding funds could thus complement each other harmoniously.

In the same way, digitization on demand could make it possible to find harmony between different actors who will each pursue their interests:

– individuals who have the possibility of seeing documents that are difficult to access and make use of digital reproduction services;

– libraries that can complete their digital libraries and provide a new service without bearing the cost;

– the patrons who can increase the status of their names by financing books on themes close to their interests;

– investors who can invest in the digitization of a book hoping that it generates enough traffic for their company or their website to benefit.

This type of service seems to meet a real need. A study on the feasibility of a digitization on demand service largely supports this conclusion [CHA 10]. Among 61 university students and 16 librarians: 91.8% of students at

Cambridge surveyed would be interested in such a service. And 65.5% of them would also be interested in the development of a print-on-demand service. According to the respondents, the correct cost of printing an item on demand would be from: £10–15 (equal to €11.81 to €17.7) for 42.9% of them and from £15 to £25 (equal to €17.71 to €29.53) for 33.3% of them; the appropriate cost of digitization on demand would have to be, for its part, from £10 to £15 (equal to €11.81 to €17.71) for 66.7% of respondents and from £14 to £25 (equal to €17.71 to €29.53) for 35% of them, while the real cost, according to [CHA 10] would be more like £40 (equal to €47.25); 44% of respondents would be willing to accept a delivery time of a week or more and only 10% estimate that this time should be only 24 h.

According to [CHA 10], the rates charged by libraries are those shown in Table 2.2.

Therefore, the cost of a digitization on demand overall remains a good deal higher than that which libraries are more accustomed to. Unlike the digitization of larger corpus or the digitization referred to as "mass digitization", with digitization on demand, it is impossible to sort documents into streams depending on their physical characteristics and it is therefore necessary to choose what material to use, configure the scanners and adjust them to each document to be digitized. This configuring time cannot be used for the digitization of several books and it becomes more difficult the use of a particular type of machine more profitable, which leads to a much larger per unit cost.

Library	Digitization on demand	Printing on demand
University of Utah	$0.05 (or €0.04) per page $20 (or €14.77) for a 400-page book	$0.05 (or €0.04) per page $20 (or €14.77) for a 400-page book
McGill Libraries (Canada)	$10 (or €6.94) fixed price for a PDF (digitization with Kirtas)	$29 (or €20.14) fixed price for a book printed with Espresso Book Machine
National Library of Australia		€8.90 per fifty-page section €35.60 for a 400-page book
National Archives	£3.50 (or €4.23) for the majority of documents	
Cambridge University Library	After 1900 and for 400 pages: £265 (or €320.58) (scan) Before 1900 and for 400 pages: £1,298.50 (or €1,570.87)	After 1900 and for 400 pages: £265 (or €320.58) (photocopy)

Table 2.2. *Rates offered by various institutions offering digitization and printing on demand*

Setting up on demand digitization services can be done without having to go through libraries and State Administrations and generating management costs for them. The individual or patron could thus order and then pay the service provider.

However, setting up this type of service could turn out to be too late. Indeed, digitization may soon become ancient history. Google Books has surpassed its initial goal and passed the threshold of 30 million digitized books. Other projects have also digitized large amounts and several thousand books are digitized each day. In 2010, Leonid Taycher, an engineer working for Google, estimated the total number of printed books produced in the world since Gutenberg at close to 130 million. According to [MUH 09], around 1 million books were published in Europe between 1500 and 1800 and five million between 1800 and 1900. Libraries now struggle to identify books that have not already been digitized. In France, at the bibliothèque Sainte-Geneviève, for example, starting from a SUDOC retrieval containing 16,000 document records at the one bibliothèque Sainte-Geneviève, only 400 remained after the elimination of doubles, off-prints, certain monographs in several volumes described sometimes as several records, sometimes as a single record, documents already digitized, documents of no interest. For its part, Google has already slowed the pace of its digitization. The market for digitization seems to be gradually reaching its limits and shrinking.

Under these conditions, it becomes more and more difficult for the heads of digitization projects to identify the documents that still deserve to be digitized and the market for a digitization on demand through crowdfunding service could also be narrowing. However, these new conditions could also, conversely, be perceived as an opportunity to make this type of service even more relevant. If mass digitization has reached its limits and can no longer be maintained, only "specialty" or "niche" digitization, which is per unit and thus on demand, would have any interest, since it makes it possible to digitize documents that are not copyrighted, following the logic of collections, large digitization programs or documents in rare languages or about very specific subjects and, once this is done, make it possible to better satisfy user needs.

Preexisting general crowdfunding platforms could also be used by libraries in order to get Internet users or patrons to finance digitization of their books without having to develop specific platforms.

The *Gold Open Access model*, used by the RevealDigital (Lyrasis) project, could also be used to obtain a return on investment on digitizations already carried out. Digitized documents would only be accessible to subscribers through *pay per view* or in libraries, but could be "freed" through a subscription to crowdfunding and sponsorship on demand. But there is no guarantee that this kind of model can still compete with immense digital libraries whose content is accessible for free.

For a more comprehensive overview, we could have also mentioned the projects Numalire, FeniXX, unglue.it, Maine Shared Collections Strategy (MSCS – maineinfonet.org), and the International Amateur Scanning League.

2.3. Printing on demand (POD): the Espresso Book Machine

Digitization on demand and printing on demand follow very similar logic. Although printing on demand is not, strictly speaking, a form of crowdsourcing, unlike digitization on demand, it is impossible to talk about digitization on demand separately from printing on demand first of all because digitization services on demand usually also offer printing on demand, second because, historically, printing on demand has sometimes preceded digitization on demand, and finally, since the "on demand" economic model is the same. Thus, instead of converting, through digitization, an object in a printed format into an electronic document following the demand of a user, we would, on the contrary, and always following the demand of a user, retro convert, through printing, an electronic document into a new document in a printed format and thus, "revive" the print original.

For several years, the number of print-runs of books has been decreasing in the publishing sector. In 2002, to adapt to this situation, the POD model was introduced. It involves printing books using a lean supply chain with inkjet printers rather than *offset* machines as was the case before, in quasi-real time, and according to consumer demand. In this way, the latter directly influences production. *POD* therefore makes it possible:

– to no longer overproduce and have too many unsold items that represent a dead loss for businesses;

– to no longer have to administrate, manage and maintain stocks that can be expensive to store and warehouse;

– to limit the costs related to the logistics of the book supply chain and in particular involving transportation;

– to no longer have to anticipate and predict the number of copies that might be sold and as a result be able to take more risks;

– to produce print-runs as close as possible to existing needs;

– to make it possible to publish works meant for very small communities, very specialized books with smaller print-runs;

– to bypass the problem of out-of-print works [BLU 16];

– to better satisfy the needs of populations reading in various languages in societies that have become multicultural.

POD is the application, in the fields of printing and publishing, of the *just in time* economic model. Traditionally, publishers, like libraries, base their modes of operation on a very different model, the *just in case* model, which consists of producing and stocking to anticipate consumption and demand or, of buying a book in case a reader needs it one day.

As we have succinctly summed up in our chronology of crowdsourcing, economic model known as *just in time* began in Japan in the 1950s and developed notably at the company Toyota. Since the space available in businesses and shops in Japan was very limited due to the constraints of geography and Japanese town planning, it was not possible to have access to a large stock of pieces of the same merchandise and it was therefore necessary to find ways to quickly replace the merchandise sold without access to large stocks. The model was subsequently largely developed and conceptualized by the Toyota Production System. It involves above all decreasing costs, avoiding unsold items and stocks of merchandise likely to gradually lose its value. With this model, supply is more directly determined by demand, production is driven, through a lean supply flow, by consumption.

In 2007, the production of printed books, via *offset machines*, increased by only 1% while that of POD books had multiplied by six between 2006 and 2007. Over the 2002–2007 period, the number of titles produced with *offset* had increased by 29% while it increased by 313% for *POD* [DOU 09]. One year later, in 2008, according to Bowker, the production of printed books in the United States with the traditional model experienced an increase of 3% while production in the form of printing on demand increased by 132%. In the United States, the number of books printed using *POD* is now higher than other methods, because of self-publishing in particular.

The cost of production with the inkjet printers that are used for the *POD* model remains higher compared to the traditional model, which uses *offset machines*. As a result, a book is 20–30% more expensive, according to Luc Spooren. Despite this inconvenience, 30,000 copies could be printed in only 2 days using the *POD* method while the same amount of printing would require 2 weeks to be produced with traditional methods [DOU 09].

After having encountered strong interest from the Government Printing Office, Internet Archive and Google Books, this economic model would necessarily also touch upon the world of digitization of libraries. Because of printing on demand, texts that have fallen into the public domain and which have been digitized and can be "revived" it in printed form, in the form of facsimiles. But this "on demand" mode of operation is very different from the traditional mode of operation of libraries and follows the *just in case* model. They purchase books and build collections in anticipation of on demand. However, with printing on demand, as with digitization on demand, there is a diametrically opposed model that applies, that of *just in time*. Its application thus would directly challenge the acquisition policies of libraries by professionals.

Thus, Lewis [LEW 10] even ventures so far as to imagine a traditional library that currently purchases 10,000 books per year. Each title costs an average of $35 to buy, $25 for its ordering and its cataloguing and $40 for its installation, storage and circulation, which is $100 per book, or 1,000,000 dollars per year for 50,000 consultations or loans per year, the use of funds involves only a minor part of the documents acquired. While, like the majority of libraries, its readership is decreasing dramatically, instead of maintaining most of its financial and human resources to buy, catalogue, equip and preserve books that are consulted less and of which only a minority will one day be consulted, could not the library in question produce, via an Espresso Book Machine, a book when a user needs one instead, and thus free the mass of salaried employees to dedicate themselves to new, more innovative, more useful and more enriching (open archives, bibliometrics, monitoring, training, digitization, *Text mining*, etc.) missions? Under these circumstances, it would cost $60,000 per year to rent an Espresso Book Machine, $40,000 per year to pay an operator (unless the library is able to retrain one of its employees) and the production of each printed book would amount to $3 while the publisher's rights would be $15 on average per printed book. Under these conditions, with an identical annual budget of 1 million dollars, the library could produce close to 40,000 books per year [LEW 10], which is four times more than before. It could have a collection built by its users and be sure that each one of the books that it conserve has been consulted at least once. The wait time for the user before having access to the book would not be more than 5 min.

Created in 2006, sold by the company OnDemandBooks and originally distributed by Xerox, the Espresso Book Machine is the result of integrating a copier, a printer and a paper cutter-folder and binder in a single machine. This photocopier with a module capable of adding a softcover binding cover to the book makes it possible to print books from the 11.4 × 12.7 cm to the 21 × 27.3 cm format.

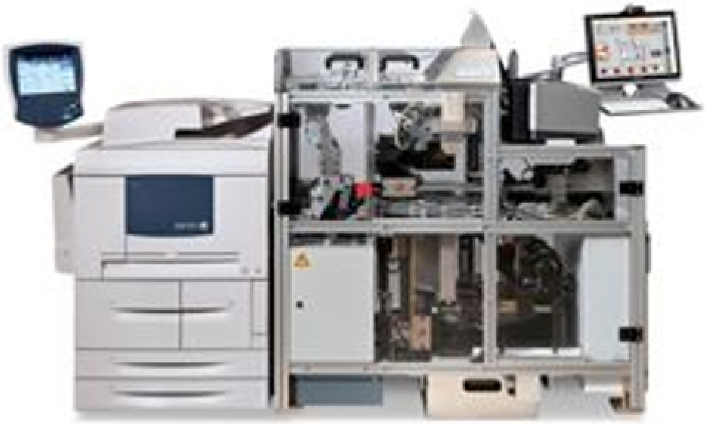


Figure 2.8. *Photograph of an Espresso Book Machine, from ondemandbooks.com. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Installed in bookstores, libraries, train stations, airports or highly frequented places, the Espresso Book Machine makes it possible to purchase *in situ*, a printed version of electronic books available on EspressNet, which offers more than eight million titles, a part of which are from Google Books (a million books), archive.org (more than 2 million books), HathiTrust, Lightning Source and Gallica. This machine also lets its users print their own productions. This type of demand is moreover very often in the majority according to feedback. Libraries can also integrate digital libraries into the Espresso Book Machine catalogue.

For a 300-page book, the time period would be 5 min according to [AND 10], but for more complex documents, this period of time could be up to 20 min. According to [DOU 09], the cost would be around $10 (equal to €7.36). According to [GEI 11], this price varies more precisely from $6 for a book of 150 pages long maximum to $10 for a book 151–450 pages long. According to [CHA 10], a 400-page book would cost $8 (€9.44) and it would be necessary to print more than 1,000 per year for the operation to be profitable. According to [WIL 11], this cost would be from $0.01 per page and, for the University of Michigan, the price for a hardcover copy would be on average $39.95 with a delivery and handling charges of $7 for the United States and $15 for the rest of the world. As for Blackwell's Bookshop in London, it offers self-publishing at £35 for the first copy. Extra books cost 5 cents per page with a minimum of £5 per book. For a 300-page book, the cost would therefore be from £15, as long as a preliminary test book has already been created. These costs are relatively low when compared to those of interlibrary loan services, which are close to close to $30. On the other hand, purchasing one of these machines is somewhat expensive and requires technical maintenance (possible paper jams, paper, ink, glue, boxes of covers, etc.). A location, a franchise or a public service delegation could be appropriate frameworks. According to [WIL 11], the cost of installation would be around $92,000 (equal to almost €68,000). The same author shows that it would be necessary to print roughly 60,000 books each year for the cost of the copy to be enough. This price could nevertheless decrease. Thus, the University of Toronto announced it had bought an Asquith press for less than €46,000. In libraries, the first Espresso Book Machine was bought by the New York Public Library on June 21, 2007. In July 2012, machines were installed in multiple libraries and bookstores in the United States (27), Canada (12), England (2) and Australia (2).

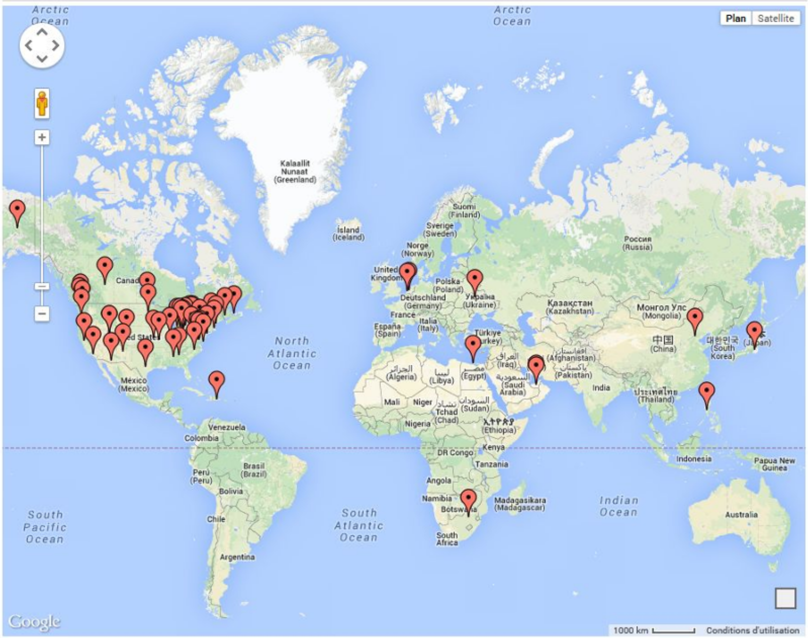


Figure 2.9. *Distribution of EBM throughout the world, according to http://www.ondemandbooks.com/ebm_locations.php (consulted on July 9, 2014). For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

This machine could offer new services to library users all the while being a source of income especially as part of a partnership with a private company such as a bookstore, for example.

Among the other uses of the Espresso Book Machine are:

– self-publishing (of books, conferences, memoirs or theses);

– the possibility of completing institutional collections (replacing missing copies or increasing copies of in-demand documents);

– interlibrary loan [GEI 11].

Interlibrary loan of a book adds up to $30 while printing on demand of a book via an Espresso Book Machine on average totals only $10, according to a study at the Virginia Polytechnic Institute and State University reported by Dougherty [DOU 09].

However, by taking this course, libraries stop being simply libraries and also become booksellers and self-publishing centers. The success of the operation will depend above all on the public that they already able to attract.

Documents digitized by libraries could therefore be sold, in the form of items printed on demand, through mail order, by businesses, or on site via an Espresso Book Machine, as part of public service delegations. In this way, libraries would trust this new public service mission to a delegate paid by the use of this service without having to ensure its management. Thus, all while offering new services to their users without having to bear the cost, libraries could even benefit from a return on investment and participate in creating economic activity. They could also increase the visibility of their digital libraries by making their content available on the sites of online bookstores, but also complete their collections printed using *POD*, to multiply the number of copies of certain frequently requested titles, replace missing copies, offer self-publishing services for memoirs, theses or any other type of writing, editing books of events for public or private institutions and to modernize their interlibrary loan services that could at the same time become less expensive for users.

The possible development of this type of service in libraries also modifies their definition and scope. The border between library and bookstore could be called into question with this economic model. As Arlitsch [ARL 11] reports, a 1979 study done by Allen Kent (*Use of Library Materials*: *The University of Pittsburgh Study*, M. Dekker, New York), a book purchased by a university library has less than one chance in two of being consulted one day. As this author suggests, it is very likely always the case. *POD* could call into question the traditional *use it or lose it* model on which the acquisition policies of libraries are founded and which is based on the anticipation of the needs of their readers by substituting an on-demand model that is more centered on the user.

Thus, just as digitization on demand made it possible to create digital libraries whose collection management policy and creation of digital collections are the work of Internet users themselves, with printing on demand, physical libraries of documents printed on reader demand could be built, collections which would be directly the work of users.

For a more comprehensive overview of the projects that have experimented with printing on demand of digitized books, we could have also mentioned the Electronic Library (eLib) and Higher Education Resources ON Demand (HERON), Amazon BookSurge (CreateSpace) and the possibilities offered by companies such as lulu.com, Lightning source, Virtual Bookworm, Wingspan press, iUniverse and Xlibris.

2.4. Participative OCR correction and participative transcription of manuscripts

Digitizing a page of a book will generate a simple photograph of the page. Starting from this simple digital image, it is impossible to search (“full text”) for a word in the document and to have its content indexed by search engines. It is also impossible to copy and paste a paragraph or generate EPUB files to be read on tablets and e-book readers. To make these things possible, the image of the text needs to be OCRized i.e. undergo OCR processing with the help of a dedicated software program. This software will determine the areas of text, the columns, tables and images (segmentation), then seek to identify which character corresponds to the image of which character. At the end of the process, the software will have produced a text file based on the image file, by identifying each of its characters, as if someone had been tasked with entering it using a keyboard.

Unfortunately, this type of character recognition processing still sometimes generates numerous errors. The quality of the OCR will depend on the quality of the digitization like the quality of the printed text. At the level of the original document, the following elements, for example, can be the source of errors:

– in the original copy: hole, discoloration, stain, fold, distortion, disparity, etc.;

– manuscript annotations;

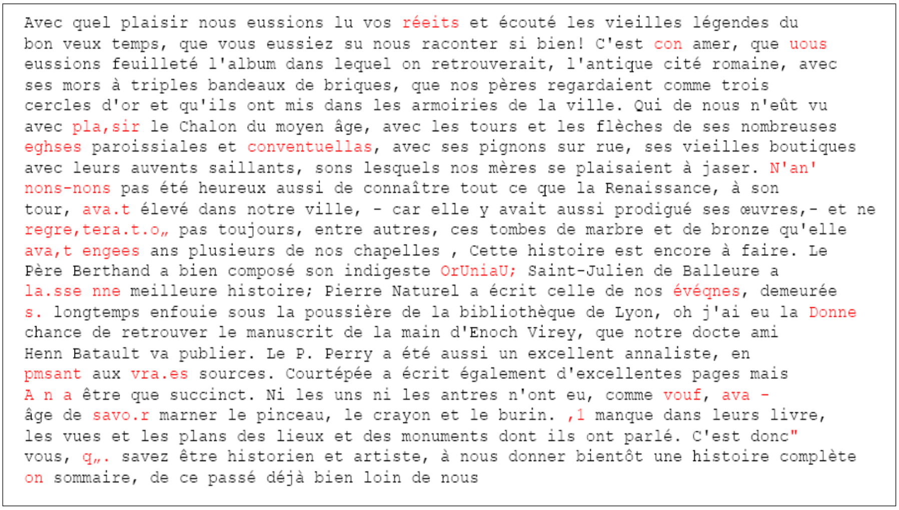
– typography: irregular (incunables, for example), badly printed, forgotten, or very unusual typographies.

Some examples of errors in interpretation are given in Table 2.3.

Printed character	Common interpretation errors by OCR software
H	li
M	in
Museum	inuseuim
Théologie	tliéologie

Table 2.3. *Examples of OCRization*

Figure 2.10 is an example of raw OCR text.

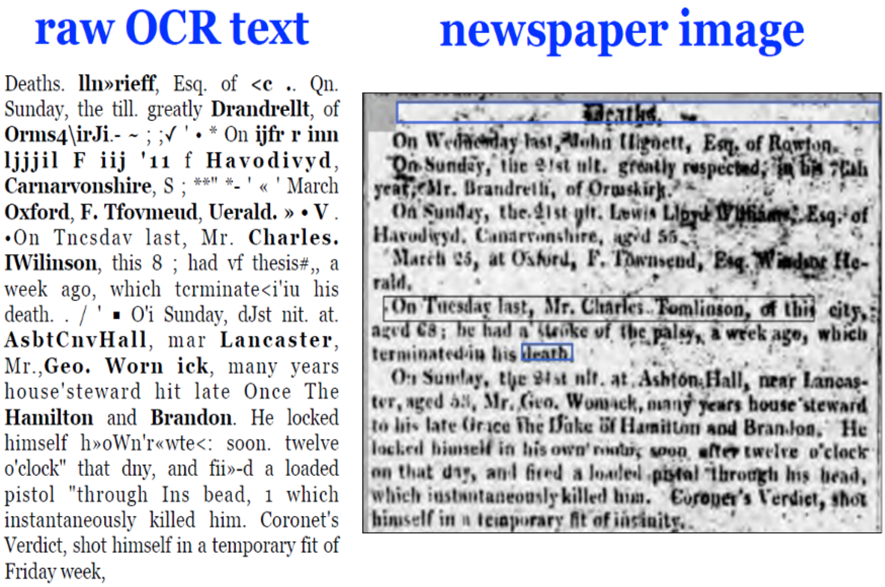


```
Avec quel plaisir nous eussions lu vos réeits et écouté les vieilles légendes du
bon veux temps, que vous eussiez su nous raconter si bien! C'est con amer, que uous
eussions feuilleté l'album dans lequel on retrouverait, l'antique cité romaine, avec
ses mors à triples bandeaux de briques, que nos pères regardaient comme trois
cercles d'or et qu'ils ont mis dans les armoiries de la ville. Qui de nous n'eût vu
avec pla,sir le Chalon du moyen âge, avec les tours et les flèches de ses nombreuses
eghses paroissiales et conventuellas, avec ses pignons sur rue, ses vieilles boutiques
avec leurs auvents saillants, sons lesquels nos mères se plaisaient à jaser. N'an'
nons-nons pas été heureux aussi de connaître tout ce que la Renaissance, à son
tour, ava.t élevé dans notre ville, - car elle y avait aussi prodigué ses œuvres,- et ne
regre,tera.t.o„ pas toujours, entre autres, ces tombes de marbre et de bronze qu'elle
ava,t engees ans plusieurs de nos chapelles , Cette histoire est encore à faire. Le
Père Berthand a bien composé son indigeste OrUniaU; Saint-Julien de Balleure a
la.sse nne meilleure histoire; Pierre Naturel a écrit celle de nos évéqnes, demeurée
s. longtemps enfouie sous la poussière de la bibliothèque de Lyon, oh j'ai eu la Donne
chance de retrouver le manuscrit de la main d'Enoch Virey, que notre docte ami
Henn Batault va publier. Le P. Perry a été aussi un excellent annaliste, en
pmsant aux vra.es sources. Courtépée a écrit également d'excellentes pages mais
A n a être que succinct. Ni les uns ni les antres n'ont eu, comme vouf, ava -
âge de savo.r marner le pinceau, le crayon et le burin. ,1 manque dans leurs livre,
les vues et les plans des lieux et des monuments dont ils ont parlé. C'est donc"
vous, q„. savez être historien et artiste, à nous donner bientôt une histoire complète
on sommaire, de ce passé déjà bien loin de nous
```

Figure 2.10. *Screen capture of a raw OCR text. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Depending on the quality of the original document and the performance of the digitization, the text obtained will include more or fewer errors in interpretation. Here, for example, is the result of OCR on an original of even lower quality from the digital library of Australian newspapers TROVE.

According to [CON 09], reporting the experiment of the British Library on 19th Century newspapers, an average of 20% of the text page on a page is not correctly OCRized. As Holley [HOL 09d] also states, the rate of OCR can vary from 71% to 98.02% from one digitized periodical to another. It is possible to improve this finding the best quality original printed copies by increasing the resolution of the digitization, by using conservation formats (TIFF or JPEG 2000), by using gray scale or in color files, by subjecting the horizontality of the lines of text or the geometry of the pages to various types of processing. Next, a large number of errors could be corrected by referring to dictionaries of words. Non-automated control and human correction will still sometimes be necessary.



raw OCR text

Deaths. **lln»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 ; **" *- ' « ' March **Oxford**, **F. Tfovmeud**, **Uerald. » • V** . •On Tncsdav last, Mr. **Charles. IWilinson**, this 8 ; had vf thesis#,, a week ago, which terminate<i'iu 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.

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

On Sunday, the 21st ult. greatly respected, in his 7?th year, Mr. Brandreth, of Ormskirk.

On Sunday, the 21st ult. Lewis Lloyd Williams, Esq. of Havodwyd, 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.

Figure 2.11. *Screen capture of a digitized newspaper and its OCR. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

When it comes to handwritten text in particular, character recognition is in its infancy [BRO 12]. Raw OCR, that is to say OCR not corrected by a human, can make reading a text on a tablet difficult or sometimes even impossible, as is interrogation by full-text searches, indexing by search engines or annotation by Text mining.

For all of these reasons, libraries with sufficient financial means outsource manual OCR correction work to service providers that rely on low-cost labor in Madagascar, India or Vietnam. An alternative would be to outsource these operations to the crowd of Internet users by allowing them to correct the texts obtained in order to improve their quality and allow better full-text searches, better indexing by search engines, produce EPUB files that can be read on tablets, the reuse of data in *Linked Open Data*, and to make possible the semantic, culturomic uses or the *Text mining* of texts.

2.4.1. *Explicit crowdsourcing: volunteer correction/transcription*

2.4.1.1. *Participative and volunteer OCR correction: the Australian newspapers digitization program (TROVE)*

Begun in March 2007 and launched in August 2008 by the National Library of Australia, this project is one of the first and most important participative OCR correction projects done by a library. It offers correction of all kinds of documents, but it is the newspaper portion that attracts the largest audience and most contributions.

There is a large amount of statistical data available in the literature that demonstrates the success of the project, one of the best benchmarks in the field. In November 2009, 8.4 million articles totaling 830,000 pages had been put online. Over 4 months, between January and May 2010, the site generated web traffic of 987,147 unique visitors. On February 8, 2013, the site includes 83,152 user accounts, 8,186 of which were active. In May 2014, a total of 129,046,297 lines were corrected because of the voluntary work of Internet users.

Thus, on average, 2,682,119 lines of text are corrected each month by close to 30,000 volunteers, taking the average of the first five months of 2014 [ZAR 14]. Ayres [AYR 13] estimates the value of 100 million lines of text corrected at 425,000 h of volunteer work, 270 years of work and 12 million euros. Based on the average costs of OCR correction for service providers of $0.50 for 1,000 characters and an average of 40 characters per line, in 2012 Brian Geiger assessed the earnings, or rather the money not spent, for Trove (68,908,757 lines corrected) at $1,378,175 [GEI 12]. In May 2014, we can assess this cost at $2,580,926 by using the following method of calculation: 129,046,297/(1,000/40) × 0.5. The calculation was confirmed by Zarndt [ZAR 14].

We can see the statistics of the number of lines corrected in Figure 2.12.

It seems that a threshold, nevertheless, has been reached and that the increase in the number of contributions is no longer proportional to that of the content put on line. The “market” of participative OCR correction might therefore have reached its limits according to [AYR 13].

Furthermore, according to [HOL 09b], 29% of the work was carried out by the 10 largest contributors who were able to devote close to 40 h per week to the work. More recently, Paul Hagon, the senior web designer at the National Library of Australia, shows that 43% of corrections (41 million lines of text) were carried out by the 100 biggest contributors to the project. The same observation was also made for the project's *tagging* activity, since 57% of tags were added by only 10 *super taggers* [HOL 10b].

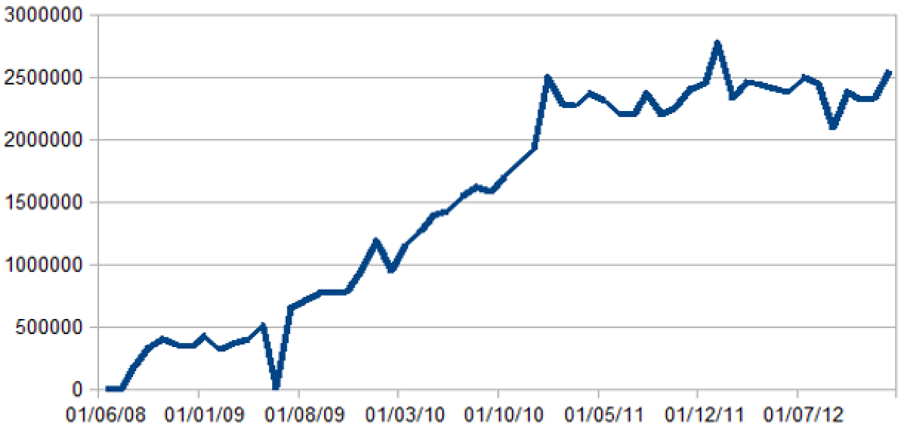
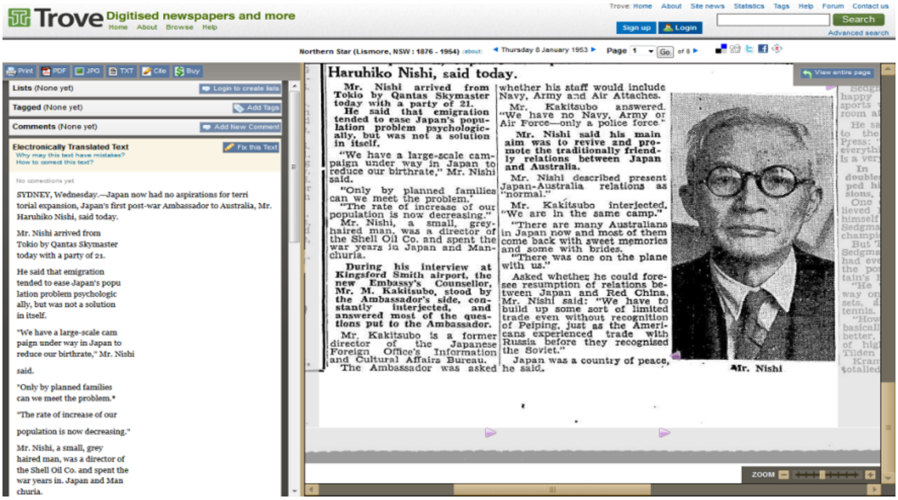


Figure 2.12. *Change in the number of corrections on lines on TROVE according to statistics obtained from the site itself (source: http://trove.nla.gov.au/system/stats?env=prod)*

As Holley [HOL 09b] shows, at the beginning of the project half of the contributions were the work of anonymous volunteers the other half was only the work of identified volunteers. However, 6 months later, 80% of the contributions were now the work of Internet users with a login. This statistic is confirmed by Paul Hagon, who puts the proportion of corrections carried out by registered users at 85%. Rose Holley estimates that this can be explained by the fact that Internet users need their contributions to be recognized and they need to be named. According to [ALA 12], the community of volunteers on Trove was, however, also interested due to intrinsic motivations (personal research, altruism, entertainment) rather than extrinsic ones (recognition, compensation).

New curation functions allow Internet users to add their own digitized newspapers and to create their own public or private collections or of documents; 40,000 collections (private or public) were created this way by Internet users, according to [AYR 13].


Trove Digitised newspapers and more
Home About Browse Help
Trove: Home About Site news Statistics Tags Help Forum Contact us
Sign up Login
Search
Advanced search
Northern Star (Lismore, NSW : 1876 - 1954) (about)
Thursday 8 January 1953
Page 1 Go of 8
Print PDF JPG TXT Cite Buy
View entire page
ZOOM
Mr. Nishi

Lists (None yet) Login to create lists

Tagged (None yet) Add Tags

Comments (None yet) Add New Comment

Electronically Translated Text Fix this Text

Why may this text have mistakes?
How to correct this text?

No corrections yet

SYDNEY, Wednesday.—Japan now had no aspirations for terri torial expansion, Japan's first post-war Ambassador to Australia, Mr. Haruhiko Nishi, said today.

Mr. Nishi arrived from Tokio by Qantas Skymaster today with a party of 21.

He said that emigration tended to ease Japan's popu lation problem psychologic ally, but was not a solution in itself.

"We have a large-scale cam paign under way in Japan to reduce our birthrate," Mr. Nishi

said.

"Only by planned families can we meet the problem.*

"The rate of increase of our

population is now decreasing."

Mr. Nishi, a small, grey haired man, was a director of the Shell Oil Co. and spent the war years in. Japan and Man churia.

Haruhiko Nishi, said today.

Mr. Nishi arrived from Tokio by Qantas Skymaster today with a party of 21.

He said that emigration tended to ease Japan's population problem psychologically, but was not a solution in itself.

"We have a large-scale campaign under way in Japan to reduce our birthrate," Mr. Nishi said.

"Only by planned families can we meet the problem.'

"The rate of increase of our population is now decreasing."

Mr. Nishi, a small, grey-haired man, was a director of the Shell Oil Co. and spent the war years in Japan and Manchuria.

During his interview at Kingsford Smith airport, the new Embassy's Counsellor, Mr. M. Kakitsubo, stood by the Ambassador's side, constantly interjected, and answered most of the questions put to the Ambassador.

Mr. Kakitsubo is a former director of the Japanese Foreign Office's Information and Cultural Affairs Bureau.

The Ambassador was asked whether his staff would include Navy, Army and Air Attaches.

Mr. Kakitsubo answered. "We have no Navy, Army or Air Force—only a police force."

Mr. Nishi said his main aim was to revive and promote the traditionally friendly relations between Japan and Australia.

Mr. Nishi described present Japan-Australia relations as "normal."

Mr. Kakitsubo interjected, "We are in the same camp."

"There are many Australians in Japan now and most of them come back with sweet memories and some with brides.

"There was one on the plane with us."

Asked whether he could foresee resumption of relations between Japan and Red China, Mr. Nishi said: "We have to build up some sort of limited trade even without recognition of Peiping, just as the Americans experienced trade with Russia before they recognised the Soviet."

Japan was a country of peace, he said.

Figure 2.13. *Screen capture of TROVE*[3]*. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

The philosophy of TROVE is founded on strong confidence in the contributions of Internet users and self-regulation. The tags are not moderated and it is possible to correct a document without having an account and without being authenticated. A reCAPTCHA anti-spam system makes it possible, however, to ensure that entries from these anonymous Internet users are not the result of robots. Moreover, all modifications are recorded and can therefore be canceled by an administrator in case of malicious contributions in order to restore the previous version. Nevertheless, as Holley [HOL 09a] shows, there was no vandalism detected over the first 6 months of the project.

As [AYR 13] points out, 62% of the visits to the National Library of Australia's different sites (including the website, catalogue, library services, etc.) came from TROVE, 75% of these visits to TROVE come directly from the search engine Google and access a particular document in TROVE, without going through a search from the TROVE site. This substantiates the idea that good referencing and good visibility on the Web are much more important than the traditional functions of library science. On average, 60,000

[3] To the right of the screen, we find the original newspaper page in photo format and on the left, the OCR to be corrected.

unique visitors per day consult the site. This number is growing significantly and was at 1.8 million visits in June 2013. Visitors remain on the site for an average of 9 min (versus only 3 min on the National Library's catalogue and a single minute on its institutional site); 40 of the visitors are from outside Australia, but come from Anglophone countries, 70% are women, 65% are more than 50 years old, a significant proportion of them are more educated and wealthier than the national average. Thus, TROVE seems in the end to mostly interest retirees passionate about local history or genealogy. This poses a problem of representation of the population that the library needs to serve and poses the question of the service's future, since there is nothing to show that future generations of retirees will also be interested in the same subjects.

Finally, although we did not mention it in the chapter dedicated to digitization on demand, the National Library of Australia was also one of the first to offer a digitization on demand service [HOL 11].

2.4.1.2. *Participative and volunteer transcription of manuscripts: transcribe Bentham*

Jeremy Bentham was an English philosopher and legal expert at the end of the 18th and beginning of the 19th Centuries, considered, along with John Stuart Mill, as the founder of Utilitarianism. Starting in 1958, the Bentham project consists of publishing the works of Jeremy Bentham ("Collected works of Jeremy Bentham") preserved previously in the form of manuscripts. As for the Transcribe Bentham project, it was launched on September 8, 2010 by University College London's Bentham Project, in partnership with the UCL Centre for Digital Humanities, UCL Library Services, UCL Learning and Media Services and the University of London Computer Centre. The project received the Ars Electronica Prize in May 2011 for the Digital Communities (€5,000) category, and was financed by a grant of the Mellon Foundation.

The project has benefitted, from April 2010 and lasting a period of one year, from financing of 262,673 pounds sterling from the Arts and Humanities Research Council, as well as two full-time Research Associates charged with developing, testing, recruiting, communicating, coordinating and moderating contributions from the UCL Library staff and a consultant from the UCL's Centre for Digital Humanities.

This project is one of the few that has released its implementation costs.

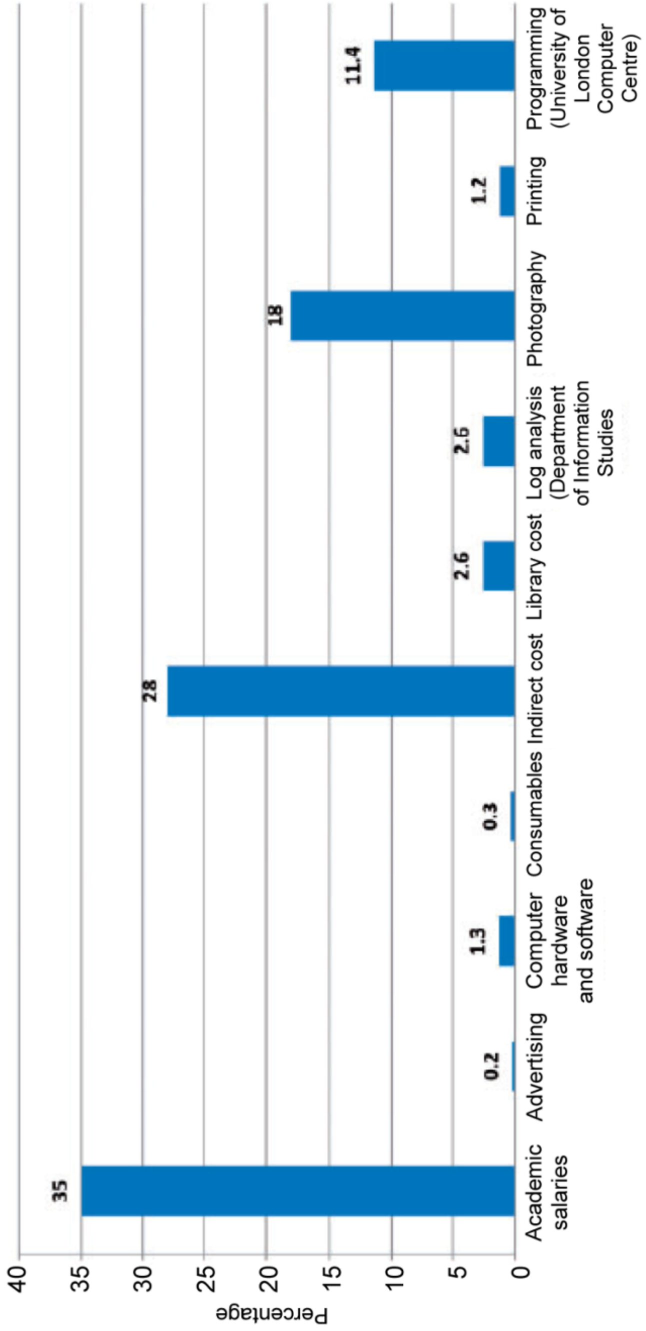


Figure 2.14. *Budget of the Transcribe Bentham project, according to [CAU 12b]*

Of the 60,000 volumes of manuscripts (30 million words) preserved at University College of London, 12,400 of them were offered this way for participative transcription and *Text Encoding Initiative* (TEI) encoding in the form of a *wiki* [MOY 11].

According to the data collected in the literature, between September 8, 2010 and March 15, 2013, 5,243 manuscripts were transcribed and there were 1,726 registered on the site on August 3, 2012.

Statistics in the form of a diagram were also published by [CAU 12b] (see Figure 2.15).

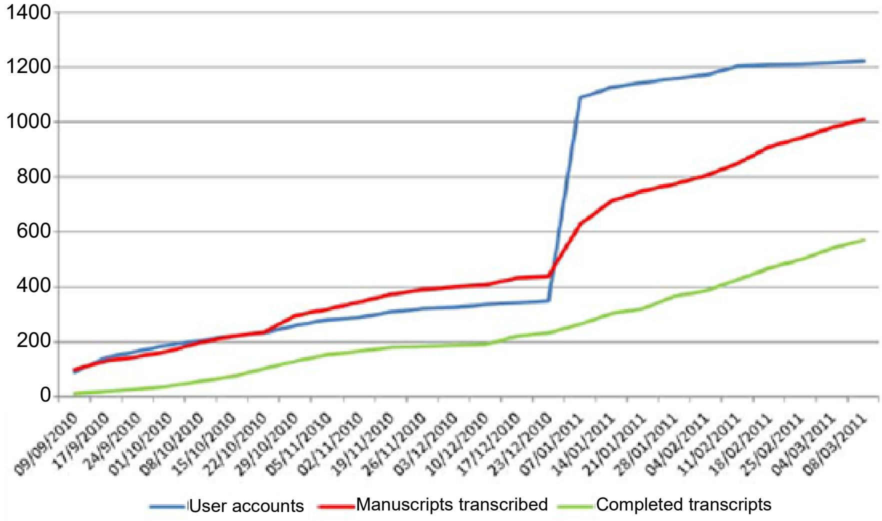
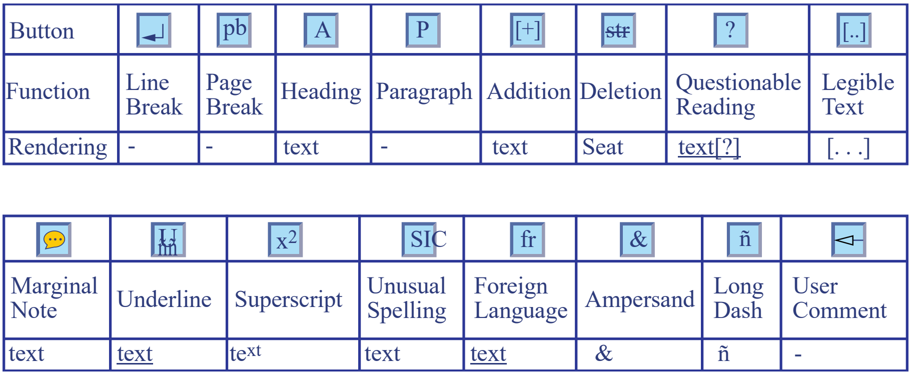


Figure 2.15. *Evolution of the number of accounts, manuscripts transcribed and completed between September 8, 2010 and March 8, 2011, according to [CAU 12b]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Thirty-two years will therefore be necessary, from May 16, 2013, to transcribe the 20,000 Bentham manuscripts. However, without the help volunteers, at a rate of 549 manuscripts transcribed per year, 131 years would have been required.

Over the pilot period, the seven biggest contributors made 70% of the contributions [MCK 12a]. Fifteen "super volunteers" each transcribed between six and 30 manuscripts [CAU 12a]. This author evokes the concept of *community sourcing* to describe the Transcribe Bentham project rather than to the extent that, following the example of numerous other projects in the cultural domain, we cannot really say that an undifferentiated mass of Internet users is contributing when it is actually a small minority of volunteers.

The MediaWiki tool, familiar to Wikipedians, was used for the transcription. The developments added within the framework of the project have given rise to a tool, which can be downloaded for free. The texts are classified by subject, date, but also by difficulty of transcription. In order to code in the TEI, without putting off beginners, the transcribers can use a toolbar in the form of a WYSIWYG interface on which each XML TEI code appears in the form of an icon.

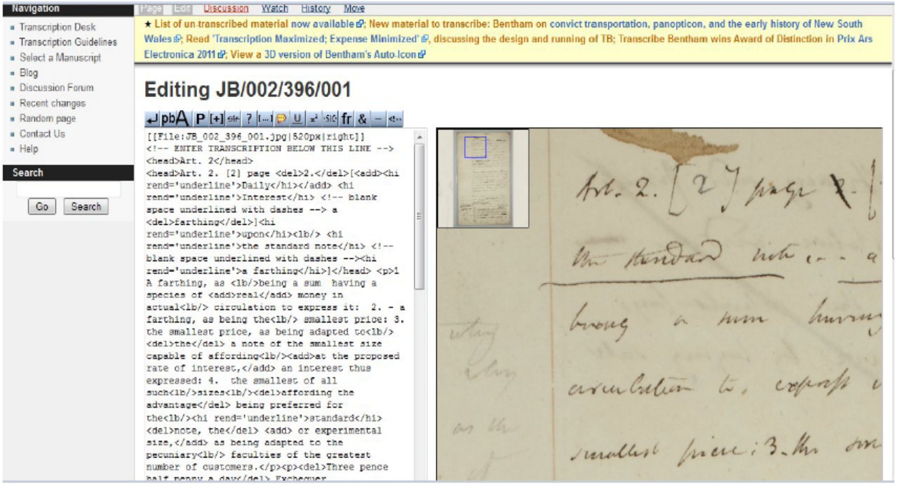


Button	↵	pb	A	P	[+]	~~str~~	?	[..]
Function	Line Break	Page Break	Heading	Paragraph	Addition	Deletion	Questionable Reading	Legible Text
Rendering	-	-	text	-	text	Seat	text[?]	[. . .]

💬	U	x^2	SIC	fr	&	ñ	◁–
Marginal Note	Underline	Superscript	Unusual Spelling	Foreign Language	Ampersand	Long Dash	User Comment
text	text	te^{xt}	text	text	&	ñ	-

Figure 2.16. *Button used by Transcribe Bentham. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

In order to contribute, authentication is required. When a transcriber validates its production, it is submitted to the editor of the project for validation by experts and then diffusion. Certain manuscripts are difficult to transcribe, and it is necessary to be pragmatic in finding a balance between the quantity of the texts transcribed and their quality. The most difficult manuscripts to transcribe, the ones written at the end of Bentham's life, are still transcribed in a more traditional way, by specialists.



★ List of un-transcribed material now available; New material to transcribe: Bentham on convict transportation, panopticon, and the early history of New South Wales; Read 'Transcription Maximized; Expense Minimized', discussing the design and running of TB; Transcribe Bentham wins Award of Distinction in Prix Ars Electronica 2011; View a 3D version of Bentham's Auto-Icon

Editing JB/002/396/001

```
[[File:JB_002_396_001.jpg|520px|right]]
<!-- ENTER TRANSCRIPTION BELOW THIS LINE -->
<head>Art. 2</head>
<head>Art. 2. [2] page <del>2.</del>[<add><hi
rend='underline'>Daily</hi></add> <hi
rend='underline'>Interest</hi> <!-- blank
space underlined with dashes --> a
<del>farthing</del>]<hi
rend='underline'>upon</hi><lb/> <hi
rend='underline'>the standard note</hi> <!--
blank space underlined with dashes --><hi
rend='underline'>a farthing</hi>]</head> <p>1
A farthing, as <lb/>being a sum  having a
species of <add>real</add> money in
actual<lb/> circulation to express it:  2. - a
farthing, as being the<lb/> smallest price: 3.
the smallest price, as being adapted to<lb/>
<del>the</del> a note of the smallest size
capable of affording<lb/><add>at the proposed
rate of interest,</add> an interest thus
expressed: 4.  the smallest of all
such<lb/>sizes<lb/><del>affording the
advantage</del> being preferred for
the<lb/><hi rend='underline'>standard</hi>
<del>note, the</del> <add> or experimental
size,</add> as being adapted to the
pecuniary<lb/> faculties of the greatest
number of customers.</p><p><del>Three pence
half penny a day</del> Exchequer
```

Figure 2.17. *The transcription interface of Transcribe Bentham, from [BRO 12]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

To train the volunteers, *screencast*[4] videos allowing them to watch online demonstrations were provided. In order to motivate them, a point system was set up with top 10 rankings, as well as a dashboard making it possible to follow the evolution of the project in real time (Benthamometer). Depending on their level, Internet users are ranked, as is done on forums, from "trainee" to "genius". The best contributors receive virtual gifts. Contributors are also thanked by name in publications of Bentham's works.

The project invested significantly in communication (press releases, radio addresses, advertisements, mailing lists, forums, social networks such as Facebook and Twitter, official blog, video[5], blogs, conferences, etc.). The publication, in particular, of an article in the *New York Times* on December 27, 2010 considerably increased web traffic. Other press releases were also distributed in the *Sunday Times*, *The Chronicle of Higher Education* (July 2010), *Deutsche Welle World Radio*, *Austria's ORF radio*. The purchase of a Google Adwords advertisement (for 60 pounds) was also tried, but without significant results. The announcement was viewed 648,995 times and led to 452 clicks, but

[4] See: http://boinc.cs.uct.ac.za/transcribe_bushman (consulted June 23, 2016, but the page is no longer accessible).

[5] See: https://youtu.be/CtEqW4WwMHU (consulted June 23, 2016).

has not led to contributions to the transcription space. A communication action to schools, universities and scholars was also attempted. Scholars were able to then participate in the project with their teacher. In total, the project's communications cost was £800 and the existence of the project was reported and mentioned in more than 70 blogs, two radio broadcasts and 13 articles in the press. On August 3, 2012, there were 853 followers on the Twitter account and 339 fans on Facebook, but these social networks seemed to have only a small impact on traffic to the site. They were therefore used more to integrate the community. In conclusion, it seems that it was actually traditional media that was the best help with recruiting.

Visitors to the site come mostly from the United States and the United Kingdom in second place. Only 6% of them have created an account [CAU 12a]. A survey was also conducted on 101 people and provided a clearer profile of the project's contributors. Unlike the TROVE project, Internet users seemed less likely to come from searches on search engines, which is probably the best justification for the communication expenses incurred during the project.

We can see in Figure 2.18 that 97% of respondents to the survey had at a least middle school education and close to a quarter had a doctorate. Close to two-thirds are women. Retirees and young degree holders are overrepresented. The overall sociology of the contributors therefore corresponds to that observed for the Australian project TROVE.

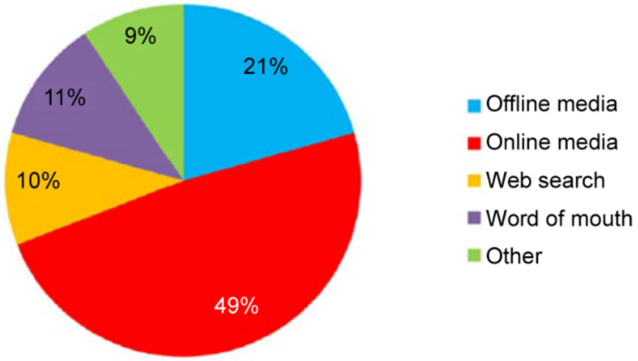


Figure 2.18. *Diagram representing how Internet users discovered the Transcribe Bentham project, according to [CAU 12a]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

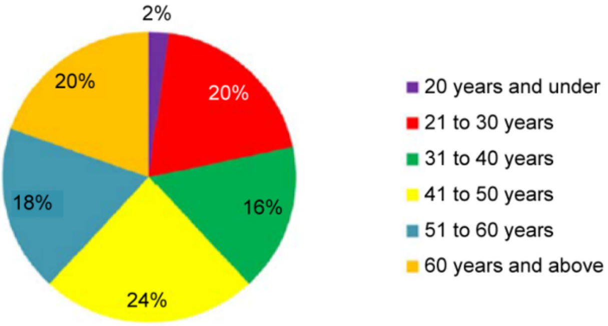


Figure 2.19. *Diagram representing the distribution of contributors to Transcribe Bentham according to age, according to [CAU 12a]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

The motivations of contributors are above all intrinsic and linked both to interest in Bentham's work and technological interest in crowdsourcing. Some volunteers also mentioned being excited by the idea of being the first to read Bentham manuscripts that had still not been edited and thus feeling as if they were doing "literary archeology".

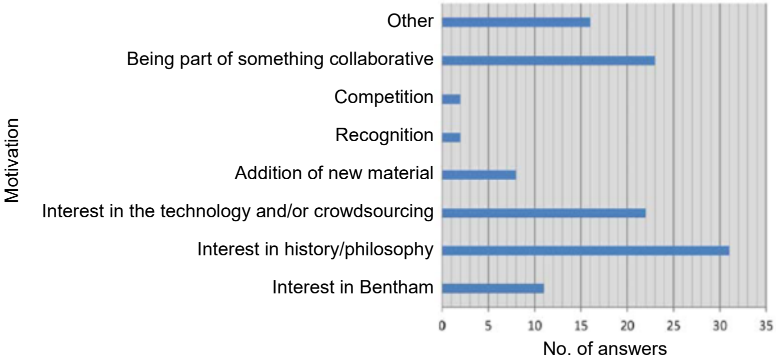


Figure 2.20. *Motivations of the volunteers of the Transcribe Bentham project, from [CAU 12a]*

For a more complete overview, we could have also mentioned the projects Distributed Proofreaders (DP or PGDP), Wikisource, California Digital Newspaper Collection (CDNC), Correct (Ozalid), Franscriptor, What's on the menu? (WOTM), Ancient Lives, ArcHIVE, What's the score (WTS), Transkribus, Do it yourself History (DIY History, Monasterium Collaborative Archive (MOM-CA).

2.4.2. *Gamification, OCR correction through play: Digitalkoot (National Library of Finland)*

Developed by the company Microtask whose slogan is “Microtask loves the work that you hate”, Digitalkoot is an application of gamification to OCR correction. Gamification consists of dividing repetitive tasks such as OCR correction into microtasks likely to be offered to Internet users in the form of games. The goal of these Internet users is therefore to enjoy themselves while contributing to a cultural project or contribute to a cultural project all while being entertained. Unlike the majority of participative OCR correction projects in the form of explicit volunteer work, OCR correction is done here out of context, without becoming aware, in a linear fashion, of the intellectual content of the document.

The project, launched on February 8, 2011, uses the platform developed by IBM, as part of the Impact project, under the name of Microtask. The Digitalkoot was inspired by *talkoot*, a very old Finnish home construction technique that is based on collective mutual assistance.

The first game available on Digitalkoot was called “Mole Hunt”. It involves hunting moles coming up out of the ground by validating as fast as possible if the images of words that they display correspond correctly to the text offered by the OCR. In this way, the game makes it possible to have raw OCR validated by comparing the players’ validations.

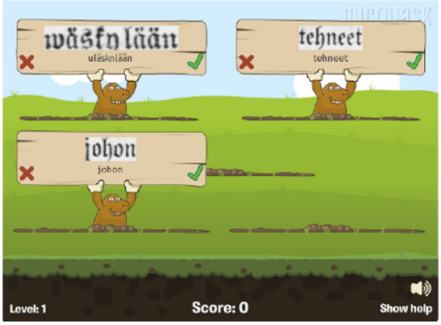


Figure 2.21. *Screen capture of the game Mole Hunt. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

A second game was offered by Digitalkoot under the name “Mole Bridge”. During this game, the Internet user must transcribe the words that are displayed in the form of images as quickly as possible. Each entry adds a new brick to the bridge that will allow the moles to cross the river without drowning. For each error, a brick in the bridge explodes. The background, speed and difficulty change with each level.

Without being aware of it, the players are evaluated during a test phase where any vandalism can be quickly detected. The overall quality of the data entered is obtained because of the comparison of the players’ transcriptions. Nevertheless, this method is expensive and requires a high rate of participation which is simultaneously difficult to obtain.

The quality of the corrected OCR obtained with the game was 99% (on only two articles containing 1,467 words for the first and 516 words for the second, 14 errors and one error were found, respectively) while the original raw OCR had an average quality of only 85%.

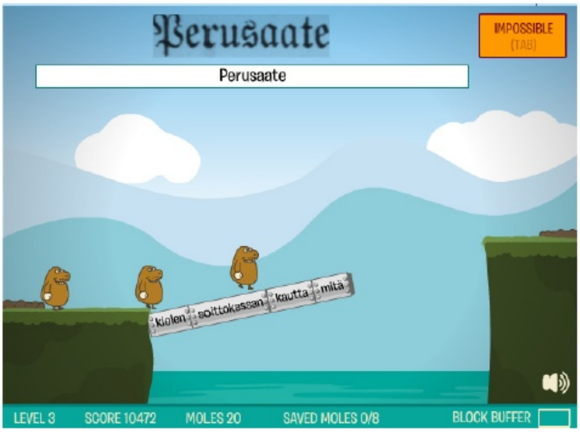


Figure 2.22. *Screen capture of the game Mole Bridge. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

In the literature, it is shown that the project generated 80,000 visitors by September 15, 2011 and that in October 2012, 109,321 contributors had accomplished 8,024,530 microtasks.

If we begin with the assumption that a page consists of an average of 220–260 words, this means that 30,000–37,000 pages were corrected by the players in October 2012. On average, we can say that they thus corrected between 154 and 182 250-page books. According to our experience leading digitization projects, the OCR correction of a page varies between €1 and €1.5. We therefore estimate that in October 2012, the project reported the equivalent of a work value of €31,000–€55,000.

The most striking result of this experiment was that almost one Finnish person in 46 has played Digitalkoot, which is 109,321 players. This high participation is the result of a communication campaign in the international press (*New York Times*, *Wired*), television and social networks having cleverly played on the patriotic feelings of the Finnish ("Start saving… Finish culture here"). The possibility of connecting using a Facebook account, which was used in 98% of cases, has also made it possible for awareness of the project to spread virally. The leaders of the project have thus estimated that one-third of the players brought their Facebook connections to the site and that the social network had attracted 99% of web traffic.

Regarding the sociology of the players, they are likely to be young (between 25 and 44 years old) compared to volunteer OCR correction projects without gamification. Women represented half of the players but carried out 54% of the work and played longer (13 min on average versus 6 min for men) [CHR 11]. However, the four largest participants were men, the best of them having played for 101 h for 75,000 transcriptions of words. This shows, once more, that, as is the case for all crowdsourcing projects, the majority of the work is the work of a small minority. Here, one-third of the work is done by 1% of the contributors.

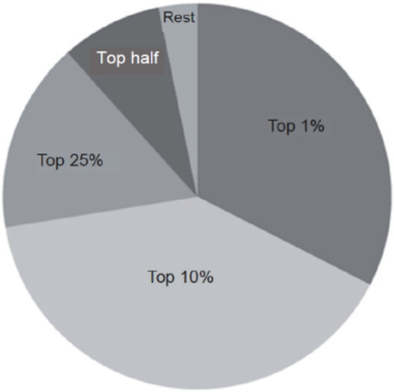


Figure 2.23. *Proportion of work carried out by 1, 10 and 25%, of the best contributors, from [CHR 11]*

The project should evolve toward the possibility of Internet users working preferentially on topics that they are interested in and toward opening up to classrooms. A new application, Kuvatalkoot, should also allow, soon, Internet users to annotate newspaper articles.

Regarding gamification, we could also have mentioned the COoperative eNgine for Correction of ExtRacted Text (CONCERT), TypeAttack, Word Soup Game, Smorball and Beanstalk from the Biodiversity Heritage Library (BHL), Tiltfactor, Metadata Games initiatives.

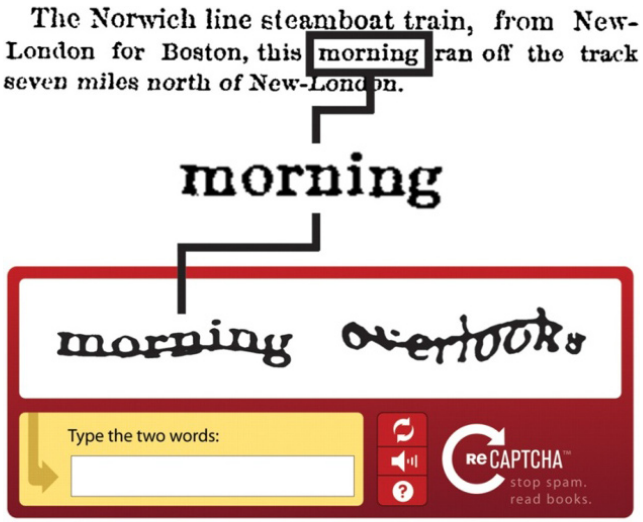
2.4.3. *Implicit crowdsourcing: involuntary OCR correction via reCAPTCHA in the service of Google Books*

CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) is a free application, developed by researchers at Carnegie-Mellon University (Luis Von Ahn, Ben Maurer, Colin McMillen, David Abraham and Manuel Blum). This application is meant to prevent malicious robots such as Googlebot from being able to submit mass requests to websites (e-mail managers such as Gmail or Yahoo, social networks, *wikis*, blogs, etc.) and paralyze servers or non-“blacklisted” e-mail accounts can be created automatically and in bulk and generate spam. We talk, for this type of application, about *Human Interactive Proof* since the system requires the Internet user to create an account and enter one or several distorted words as a way to prove that he or she is definitely a human and not a robot. This system was inspired by the Turing test. The Turing test, described by computer scientist Alan Mathison Turing in 1950, started from the hypothesis that a computer could be considered intelligent if it becomes able to distinguish the conversation of a human from that of a computer. According to the same principle, the CAPTCHA test makes it possible to tell a human from a computer, the latter being incapable of reading distorted words.

However, as explained in an article published by the journal *Science* [VON 08], reCAPTCHA has also been used, in addition to its original purpose, in order to allow correction by Internet users of OCRized text from the digitized *New York Times*. Since September 17, 2009 and its purchase by Google, this program is, currently, used as part of the Google Books program to have Internet users correct, through implicit crowdsourcing, the text of millions of books digitized by Google. Its slogan has become “Stop spam, read books”. Since March 2012, Google also uses reCAPTCHA in order to have Internet users correct photos of street numbers taken from Google Street View in order to refine Google Maps' geolocalization.

With reCAPTCHA, Internet users are not always as aware that their entries, for security reasons, are used to correct content on Google Books and Google Maps. Under these conditions, one could refer to this type of crowdsourcing as "unconscious crowdsourcing", "involuntary crowdsourcing" or "implicit crowdsourcing". Indeed, the participation of Internet users is not necessarily unconscious or involuntary either, while it remains, clearly, in every case, implicit.

Implicit crowdsourcing takes into consideration the fact that a very small minority of volunteers contribute to volunteer crowdsourcing projects (moreover it should be referred to as *community sourcing*). Taking into account the limits of this explicit crowdsourcing, the risk of saturation and difficulties in recruiting new volunteers, implicit crowdsourcing consists of using the current activities of Internet users on the Internet in a clever, ecological and economical way, and to divert them for other ends. In this way, the energy produced by Internet users in order to reenter words and create accounts on websites is recycled for the benefit of Google's digital library.

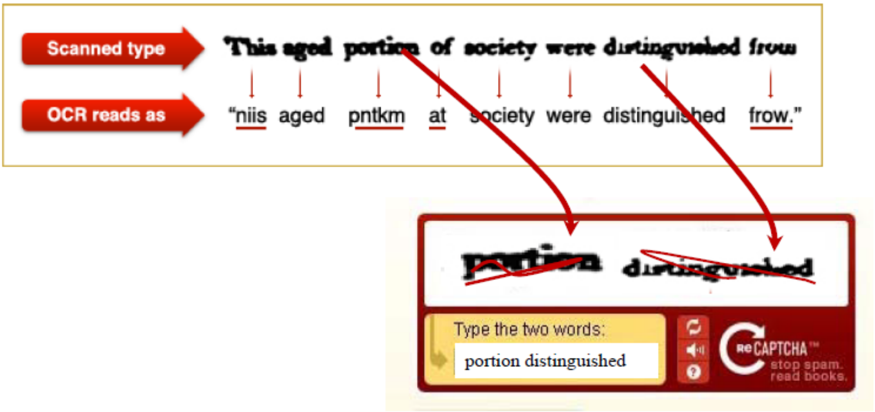


The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

Figure 2.24. *Diagram explaining how reCAPTCHA works, according to the site Google.com. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Books digitized by the Google Books project have systematically been subjected to character recognition by two different OCR software programs. The differences between the texts obtained are compared with language dictionaries. Around 25% of the OCRized words are considered likely to be recognition errors. They are then sent to reCAPTCHA to be corrected by Internet users. For their part, Internet users, in order to create accounts on websites, are required to enter two distorted words to prove they are not robots. One of the two words, already corrected and validated, serves a security purpose. The other word is used for the purposes of having Raw OCR from Google Books be corrected by Internet users.

The traditional method of comparing of entries, well known to businesses that perform correction of raw OCR is then used. The same word is sent to three different people on the Web with different distortions in order to avoid the same distortions generating the same entry errors. If the three entries from the three Internet users are exactly identical, the correction is validated. In the opposite case, the word to be corrected is offered to additional Internet users until one of the proposals totals 2.5 votes (knowing that one Internet user counts for one vote and that one of his or her proposals counts for 0.5 vote). Sometimes, Internet users can be unable to identify the word to re-enter. They can then ask that a new word be displayed. If a word is offered this way six times and flagged as illegible by Internet users each time, the system will consider it impossible to identify.


Scanned type
This aged portion of society were distinguished from
OCR reads as
"niis aged pntkm at society were distinguished frow."
portion distinguished
Type the two words:
reCAPTCHA™
stop spam.
read books.

portion distinguished

Figure 2.25. *Another diagram explaining how reCAPTCHA works, from [IPE 11]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

The leaders of the reCAPTCHA project have provided the statistics presented in Table 2.4.

Number of Internet users necessary to obtain a valid correction	Part of the total of words corrected that this situation represents (%)
Two Internet users	68
Three Internet users	18
Four Internet users	7
Five Internet users	3
Six or more	4

Table 2.4. *Statistics of the number of Internet users necessary to correct a word, after [VON 08b]*

In total, this system would make it possible to obtain, on average, a corrected OCR of 99.1%.

In 2008, reCAPTCHA was installed on 40,000 websites (including Facebook and Twitter) and made it possible to validate more than 440 million words; 17,600 books and 1.2 billion words were offered during the year 2008. According to the statistics from 2012, close to 200 million words are entered this way every day by Internet users who dedicate 12,000 h of work to it per day and this pace is increasing. In 2007–2008, more than a billion reCAPTCHA were resolved this way [CON 09].

Ipeirotis and Paritosh [IPE 11] make the following calculation. There were 40 million reCAPTCHA entered each day in 2008, which means 40,000 books corrected per day. Under these conditions, it would only take 12 years to correct 100 million books. Nevertheless, this calculation is probably mistaking 40 million entries for 40 million validations. Yes, it is necessary to compare several entries to obtain one validation. Moreover, if 40 million words equal 40,000 books, a book would only contain 1,000 words. Yet, it does not seem like Google Books is focusing on reprint requests.

In order to know the number of validated words produced each day and to assess the value produced by this work, we have therefore produced a new estimate (Table 2.5).

In order to evaluate approximately how many books are corrected this way because of Internet users, we begin with the assumption that there are, on average, 75,000 words in a 300-page book and an average of 250 words on each page. If 86 million words are corrected and validated every day, we can assume that 86,000,000/75,000 = 1,147 books are corrected every day, which is 1,147 × 30 = 34,410 books every month and 1,147 × 365 = 418,655 books per year.

During a 2011 TED conference[6], the founder of the project, Luis Von Ahn, spoke of half a million books per year. This number is somewhat close to our calculations and which seems to us more reliable that the other estimates mentioned previously.

It nevertheless seems that reCAPTCHA could be shortly abandoned by Google, which is also slowing down its digitization program. Beginning in 2015, when Google announced the launch of this project, the goal of digitizing 15 million digitized books was announced and left a good number of professionals skeptical. This goal was, nevertheless, reached and more than doubled, the bar of 30 million books having been passed in 2013. With a speed of 418,655 books corrected per year, it would require more than 70 years to correct the OCRized texts of these 30 million books. The number of books which have been printed since Gutenberg's invention of the printing press is estimated at 129,864,880 by Leonid Taycher, an engineer at Google according to an article published on its blog on August 5, 2010. Correction of the OCRized texts of these 129,864,880 books would therefore take 310 years, according to our calculations. Nevertheless, the number of Internet users and therefore the number reCAPTCHA entries may be increasing.

[6] See: https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration (consulted June 23, 2016).

Number of Internet users needed to obtain a valid correction	Portion that this situation represents out of the total of words corrected	Number of words entered per day in this situation (on the basis of 200 million words entered each day)	Number of words validated per day in this situation
Two Internet users	68%	200 million × 0.68 = 136 million	136/2 Internet users = 68 million
Three Internet users	18%	200 million × 0.18 = 36 million	36/3 Internet users = 12 million
Four Internet users	7%	200 million × 0.07 = 14 million	14/4 Internet users = 3.5 million
Five Internet users	3%	200 million × 0.03 = 6 million	6/5 Internet users = 1.2 million
Six or more	4%	200 million × 0.04 = 8 million	8/6 Internet users = 1.33 million maximum
	Total		86 million words validated each day

Table 2.5. *Statistics collected in the literature regarding the reCAPTCHA project*

But 70 years to correct the current content of Google Books or the 310 years to obtain corrected texts of every printed book ever produced are short periods of time compared with the speed of public libraries for correcting the OCR of texts. Indeed, these corrections are rarely done by libraries, which offer low-quality OCR whose content is generally impossible to consult on readers, and the most often difficult to index, to search, to extract and to reuse. When an OCR correction is carried out, libraries use service providers who use the cheap labor in Madagascar, India or Vietnam.

If Google Books did not use the reCAPTCHA system and made use of this same type of service, it would have to spend between 1 and 1.5 euros per page of corrected text, which means a 300-page book would cost from €300 to €450. This is a significant amount, which explains why libraries correct only very little of the OCR of the texts that they digitize. According to our calculations, we have estimated that Google has the OCR of around 1,147 books corrected by Internet users each day, which is 418,655 books per year. We can therefore estimate that implicit crowdsourcing saves it from paying between €344,100 (€1,147 books × €300) and €516,150 (1,147 books × €450) each day. We therefore estimate that Google benefits from 146 million euros per year of free work because of implicit crowdsourcing. This is a much larger budget than that which libraries could spend on OCR correction. In terms of working time, if we consider that employees correct at a speed of 60 words per minute, this represents the effort of more than 1,700 people working 35 h a week.

If Internet users work for free several hundreds of thousands of hours each day and, in the majority, without knowing it, for the Google Books project, this energy was, in every case, used for reasons of security, in accordance with reCAPTCHA's primary purpose. Google, aware of the value of this energy, has very cleverly and ecologically also reused it for its digitization program, a program in which everyone can, moreover, benefit. The idea of reusing reCAPTCHA in order to improve the quality of the OCRized texts remains a great example of innovation stemming from open-mindedness and transdisciplinarity. Two fields of expertise as different as computer security and the digitization of heritage have it as a common innovative application.

Nevertheless, Google was recently able to create an algorithm capable of bypassing reCAPTCHA. A new "No Captcha reCAPTCHA" mechanism with logic questions to prove that it is actually a human who has answered them would then replace the previous method.

2.4.4. *Paid crowdsourcing: the Amazon Mechanical Turk market place*

In the 18th Century, the Mechanical Turk (or automated chess player) was an automated machine that belonged to Baron von Kampelen and was supposed to be endowed with artificial intelligence allowing it to play chess. This machine would later notably beat Napoleon Bonaparte and Benjamin Franklin. In reality, the machine was finally operated by the human intelligence of a person hidden inside. Amazon was inspired by this story of a human hidden in a machine in order to demonstrate the need to use human intelligence to reach goals still impossible for machines and to show that certain work, which we think are done by machines, is, in reality, done by hidden humans. For this reason, the company Amazon launched, largely for its own needs, on November 2, 2005, the Amazon Mechanical Turk Marketplace (AMT), a paid crowdsourcing space on the Web where institutions and companies can offer microtasks known as *Human Intelligence Tasks* (HITs). Internet users come to search for tasks to carry out. Most of the time, these tasks do not require many qualifications but remain impossible to have done by algorithms or by programs. They involve, for example, transcription of video or audio recordings, indexing of documents or images, classification, summarizing, votes, identification of images, notably pornographic images, editing of comments and reviews on participative sites, adding relationships or *likes* on social networks. It can even sometimes involve correction of raw OCR texts.

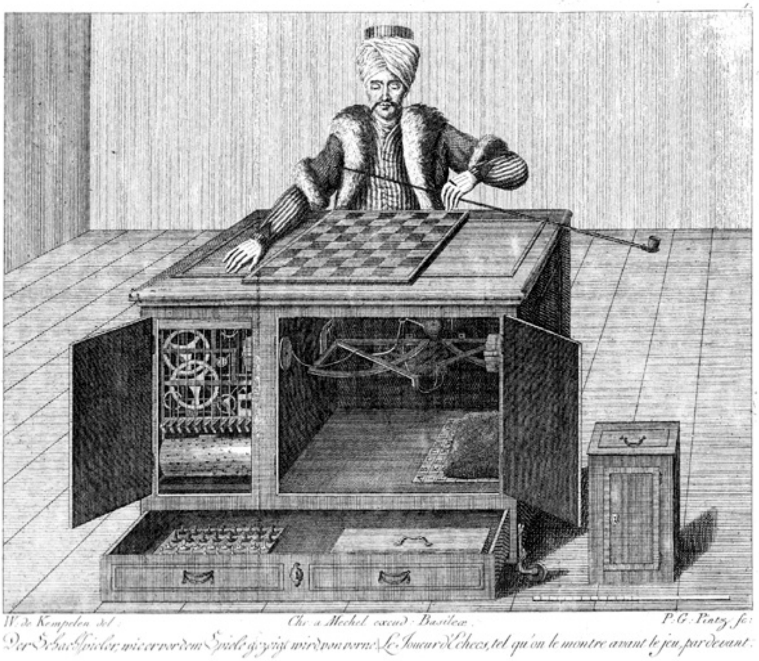

W. de Kempelen del:
Chr: a Mechel excud: Basileæ.
P.G. Pintz fc:
Der Schachspieler, wie er vor dem Spiele gezeigt wird, von vorne. Le Joueur d'Echecs, tel qu'on le montre avant le jeu, par devant.

Figure 2.26. *The Turkish chess player, Tuerkischer schachspieler windisch by Karl Gottlieb von Windisch, 1783, public domain via Wikimedia Commons*

Amazon earns money because of a commission of 10–20% for the setup and maintenance of this service to businesses and Internet users and thus has an income of between 1 and 30 million dollars per year for this service [IPE 10a]. The commission earned by Amazon increased in 2015[7].

Before hiring a worker, the sponsor can require certain qualifications and set up selection tests. Statistics regarding the workers' reputations of the workers are also accessible following the model of e-commerce sites, which make it possible to be sure about the statistics of a particular vendor. Once the work has been carried out, the sponsor can validate or refuse it.

In 2011, the platform had assembled no fewer than 500,000 workers from 190 countries. In 2014, around 200,000 HITs for a value of $40,000 were exchanged each day.

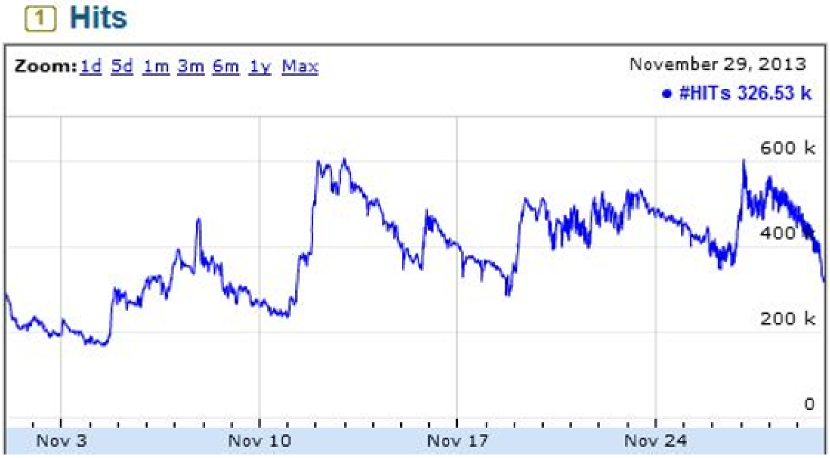


Figure 2.27. *Number of HITs in November 2013, according to the Mechanical Turk tracker*

According to [IPE 10b], workers in the *marketplace*, come mainly from the United States, were younger and more educated than the general population and were above all interested in a way to earn a little money. Their motivations are therefore simultaneously intrinsic (a profitable and fun way to spend free time) and extrinsic (a source of revenue) [KAU 11]. However, with the increase in unemployment in Western countries and since

[7] According to Nicolas Gary (24/06/2015). "Le Mechanical Turk d'Amazon augmente ses tarifs d'intermédiaire". See: actualitte.com.

the platform began allowing Indians to be paid in rupees, this type of motivation could give rise to satisfaction of much more necessary needs.

In February 2010, a study was conducted by Panos Ipeirotis on the platform by paying each participant 10 cents, then another study, conducted by [ROS 10] was also produced by also rewarding $0.10 to each worker responding to a survey on the Amazon Turk Mechanical Marketplace in less than 2 min. These surveys show that the portion of American workers (47%) will decrease in favor of Indian workers (34% for Ipeirotis, 36% for Ross). This trend is moreover relatively consistent with that reported by [FOR 11] and which shows that Indian workers represent 10% of workers on the platform in 2008, 33% in 2010 and 50% in May 2010.

American workers were more likely to be female, while the Indian workers were, on the contrary, more likely to be male, but according to [ROS 10], the proportion of men will have a tendency to increase.

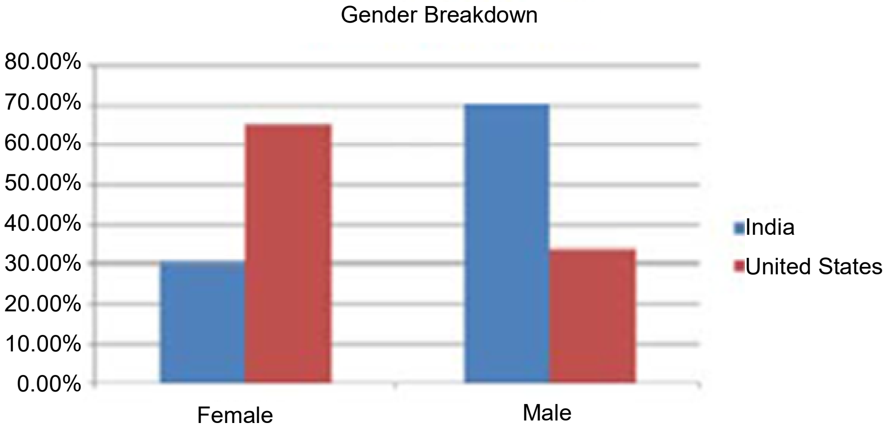


Figure 2.28. *Distribution of Indian workers and American workers on AMT by sex, according to [IPE 10b]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

If we consider that American women are more affected by unemployment and part-time work, this result is unsurprising.

The workers generally seem to be relatively young and their average age had the same tendency to decrease. They are childless and tend to be quite well educated (Figures 2.29 and 2.30).

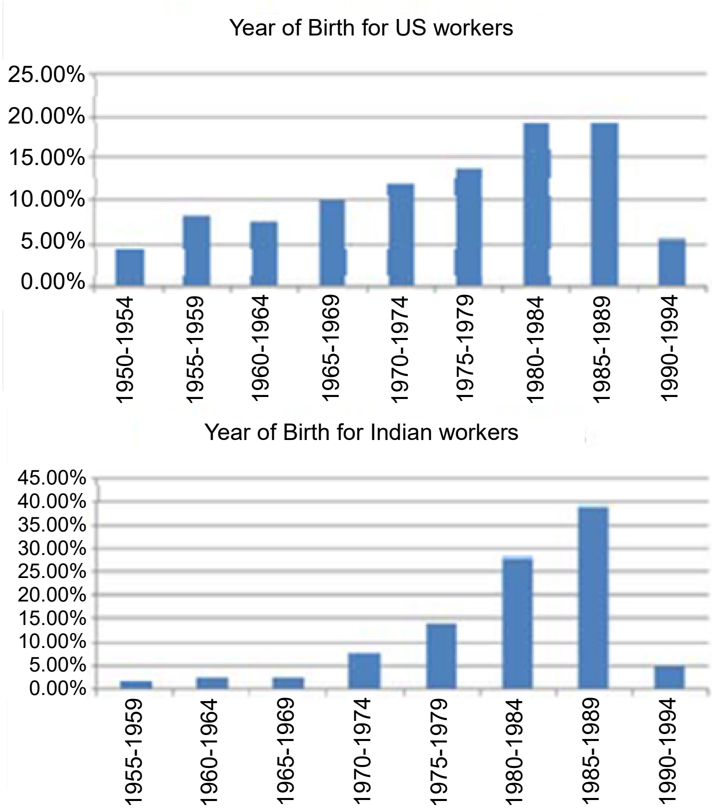


Figure 2.29. *Birth year of workers on the AMT, according to [IPE 10b]*

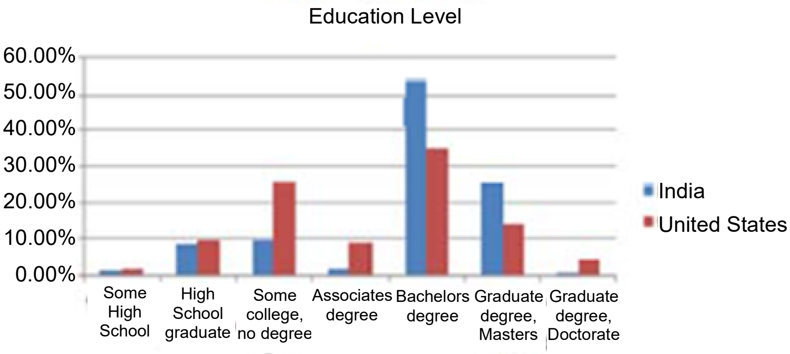


Figure 2.30. *Educational level of workers on the AMT, according to [IPE 10b]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

The majority of them spend between 4 and 8 h per week on the platform (Figure 2.31).

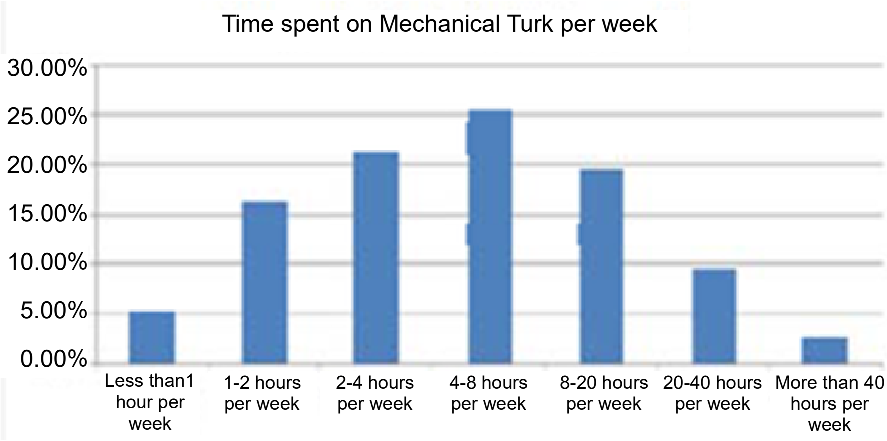


Figure 2.31. *Average time dedicated to the AMT, according to [IPE 10b]*

The large majority of workers devote less than 5 h per week to it, but all the same 18% of them worked more than 15 h per week.

They earn between $1and $5 per week (Figure 2.32).

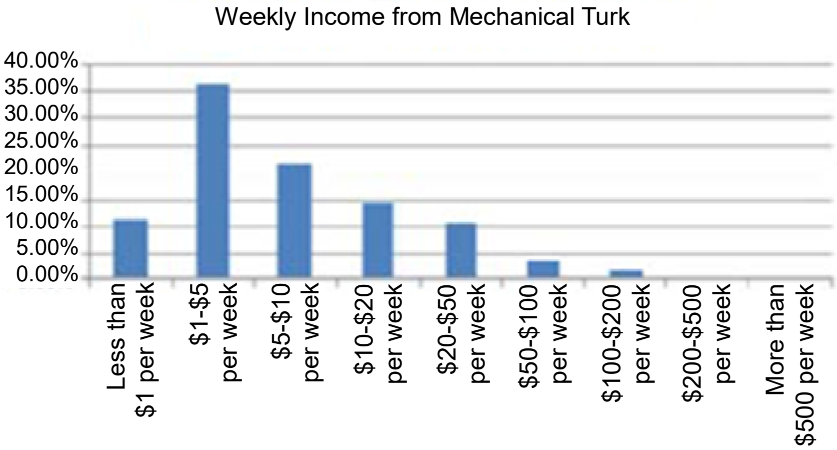


Figure 2.32. *Average income made from the AMT, according to [IPE 10b]*

The average value of HIT was only 7.9 cents [ROS 10]; 10% of HITs were paid only $0.02 maximum, 50% around $0.10 and 15% of them were paid $1 or more [IPE 10b]. Under these circumstances, and according to a survey of 400,000 workers registered on MTurk, the average American user earned $2.30 per hour in 2009 and the Indian user $1.58 per hour [ROS 10]. But few of these workers (27%) considered it to be their primary source of income (Figure 2.33).

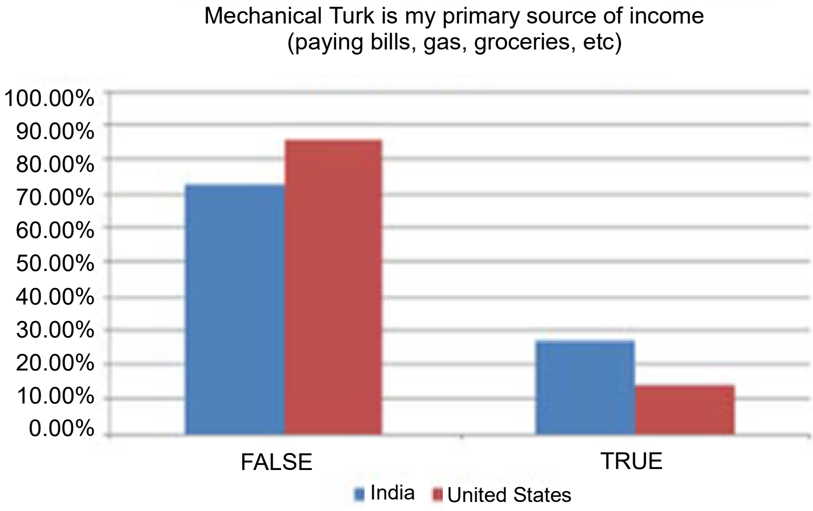


Figure 2.33. *Number of workers stating that AMT is their primary source of income, according to [IPE 10b]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Only 0.1% of businesses that have offered the most HITs represent a sizeable 30% of the activity on the platform. In the same way, according to [FOR 11], 80% of HITs were carried out by only 20% of workers who spent more than 15 h per week on the platform. These workers could number between 3,011 and 8,582 according to the study’s calculations; 20% of them consider this activity as their principle source of income and 50% only a secondary source of income.

The less dedicated workers are motivated by means of passing the time, the opposite of the most dedicated workers who would be more motivated by the variety of tasks and the identification with a community.

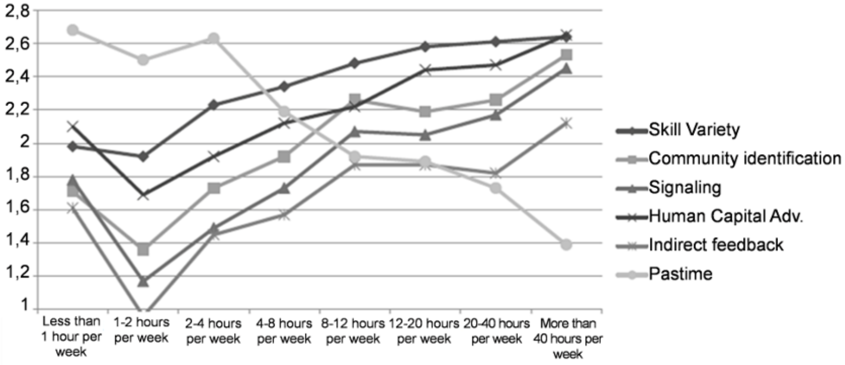


Figure 2.34. *Types of motivation according to the greater or lesser dedication of workers on the AMT platform, according to [KAU 11]*

Depending on the cumulative experience (number of approved HITs greater or equal to 1,000) and quality of the work provided (rate of approved HITs greater or equal to 95%), a worker can obtain the qualification of *Master worker*. The flexibility and autonomy provided by this type of work can be satisfying including from the point of view of salaried employees. Thus, a 28-year-old American worker testifies that based on 5,000 HITs per week, she works full time and earns more money than at her old job, all while having the freedom to choose, depending on her needs, skills and desires, her place of work, her work schedule, how long she works, the type of work and even the sponsors that have become her customers [DEN 13]. In a context of the development of "on demand" models where *jobcrafting* is expanding and where salaried workers like to modify their job descriptions more and more, this type of platform could therefore be very attractive and benefit employers as much as employees.

Nevertheless, the omnipresence of dishonest workers and unbanned and unpunished spammers on the platform leads to a depreciation in the value of work on the platform and at very low prices, the best workers end up leaving [SAG 11]. As a result, according to Ross reported by [SAG 11], 70% of Turkers have been using the platform for less than 6 months. University researchers were the first users[8].

[8] According to Nicolas Gary (24/06/2015). "Le Mechanical Turk d'Amazon augmente ses tarifs d'intermédiaire". See: actualitte.com.

Other platforms based on the same model as Amazon Mechanical Turk exist and function according to a similar model: 99design, CloudCrowd, Cloud-Flower, CrowdFlower, eLance, Foule Factory, Freelancer, Guru, Innocentive, ManPower, Mob4hire, MobileWorks, oDesk, Postmates, quora.com, Samacource, sparked.com, TaskRabbit, Topcoder, Trada, Turkit, uTest.

Regarding the use of paid crowdsourcing for cultural heritage digitization projects in libraries which interests us more specifically, experiments in the transcription of manuscripts and OCR correction are related [LAN 11, SAY 11]. An institution, for example, puts its raw OCR on Google Docs open for writing then pays Internet users around the world to carry out OCR correction via the Amazon Mechanical Turk Marketplace. The list of URLs of each Google Doc was simply placed a CSV file on the platform. After the correction work was carried out, other Internet users were then paid to control the corrected OCR in which the errors were purposefully introduced order to verify that the work had been carried out correctly.

Here, in Table 2.6, are the results of the comparison between the costs incurred by the Amazon Mechanical Turk marketplace and an estimate of what the costs of a traditional service provider would have been.

Work carried out	Costs (with the Amazon Mechanical Turk Marketplace)	Estimates (with a traditional service provider)
Transcription of a document six to eight pages long	$0.08 (equal to €0.06) in roughly one week	
Proofreading of the transcription of a document six to eight pages long	$0.10 (equal to €0.07) in roughly one week	
Transcription with quality control of a document six to eight pages long	$0.18 (equal to €0.13)	From $2 to $8 (equal to €1.45 to €5.80)
Total cost 72 pages	$22.86 (equal to €16.50)	From $144 to $576 (equal to €104.30 to €417.30)
Total cost 200 pages	$60 (equal to €43.50)	$400 (equal to €290)

Table 2.6. *Comparative costs between OCR correction via the AMT and via a service provider*

Another experiment organized around paid crowdsourcing has also been related involving the annotation of engravings of flowers preserved at the Amsterdam Rijksmuseum with the help of the platform CrowdFlower [OOS 14a, OOS 14b]. However, the crowd was perhaps not the best target for this project and the Rijksmuseum realized that what it really needed was amateurs, experts, self-taught people and retired professionals, i.e. to call upon *community sourcing* rather than crowdsourcing [DEB 12].

Taking inspiration from calculations of Ipeirotis and Paritosh [IPE 11], Geiger and Zarndt [GEI 12], and Zarndt [ZAR 14] and our own estimates, we have estimated what the financial rewards from different participative OCR correction projects would be by calculating what institutions would have paid if they had used professional labor to carry out OCR corrections.

Project	Unspent money
California Digital Newspaper Collection	$53,130 cumulatively in June 2014
TROVE	$2,580,926 cumulatively in May 2014
Digitalkoot	Between €31,000 and €55,000 cumulatively in October 2012
Google Books and reCAPTCHA	146 million euros per year at the 2008 rate

Table 2.7. *Estimate of the costs not paid for OCR correction services because of the use of crowdsourcing*

Crowdsourcing applied to OCR correction therefore presents an economic issue that is far from being insignificant. It would allow libraries to avoid wasting public funds that have become scarce, the way they are doing currently by having the work done by labor in developing countries and under sometimes unethical conditions all while making it possible to improve the quality of the texts produced as part of their digitization projects in order to offer better possibilities for full-text searches, better visibility on the Internet and better indexing of content by search engines to produce files that are readable on tablets and to allow reuse and semantic exploitation of textual data.

On reading this comparative table, it can also be seen that the results obtained by Google Books with reCAPTCHA are completely different from the best results obtained by library sites relying on traditional crowdsourcing with volunteers. The type of crowdsourcing used is also very different since in some cases, the Internet user participates voluntarily in the OCR correction (TROVE, California Digital Newspaper Collection, etc.) and in others, it is used involuntarily without even being aware of it (reCAPTCHA). This type of unconscious and involuntary crowdsourcing is generally referred to as implicit crowdsourcing.

Beyond the division between implicit and explicit crowdsourcing, the study of various projects has allowed us to differentiate projects with or without gamification, from the correction of printed texts or transcriptions of manuscripts, correction in context or out-of-context correction.

Traditional OCR correction in the context of text corresponds well to the motivations of volunteers who would like to profit from their contribution in order to learn about and read texts that interest them. As a result, we have seen some volunteers become specialists in a particular subject because of their participation in explicit crowdsourcing projects.

However, this type of crowdsourcing remains less effective than gamification or implicit crowdsourcing that offers microtasks and out-of-context correction, which do not allow for the personal development of the contributors.

As we have already stated, implicit crowdsourcing takes into account the fact that only a small minority of individuals participate in explicit crowdsourcing projects. It is therefore more efficient to recycle and reuse the energy of Internet users during their current activities on the Internet. The explicit, more traditional crowdsourcing, used for more than 10 years, could, moreover be reaching its limits.

In fact, as Ayres [AYR 13] suggests and as the statistics of the TROVE project seem to indicate, we may now have reached a critical threshold for crowdsourcing. The time available to spend on volunteer work available on the Web is not infinite and it is spread out between a larger number of projects. Implicit and explicit crowdsourcing and could thus coexist and complement each other. In the first case, Internet users play or create an account on the Web and involuntarily improve the quality of unspecified texts (or among the most consulted on the Web). In the second, they seek to simultaneously consume specific texts and to improve their quality.

Nevertheless, the improvement of the character recognition capacities of software could also render participative OCR correction projects obsolete in the next 5 years, the human eye could gradually be replaced by algorithms. The border between what it is possible to have done automatically by a machine and that which must still be performed by a human is a blurred line. As OCR technologies improve, the use of crowdsourcing will become less and less necessary and requiring volunteers to correct texts could thus soon only be useful for transcription of manuscripts.

In every case, if we consider the changes in the number of corrections in the largest participative OCR correction project, the Australian TROVE project, we can observe a clear a slowing down in recent years (Figure 2.35).

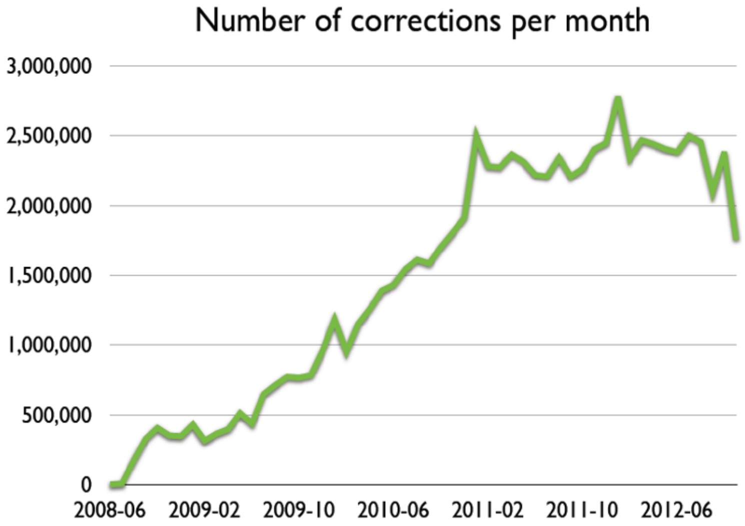


Figure 2.35. *Number of corrections on TROVE between 2008 and 2012, according to [HAG 13]*

This stagnation cannot be explained by a stagnation of the amount of content offered for correction, since it has continued to grow as Figure 2.37 illustrates.

In theory, explicit crowdsourcing is addressed to the crowds of participants. In reality, the majority of contributions come from a very small minority of motivated volunteers.

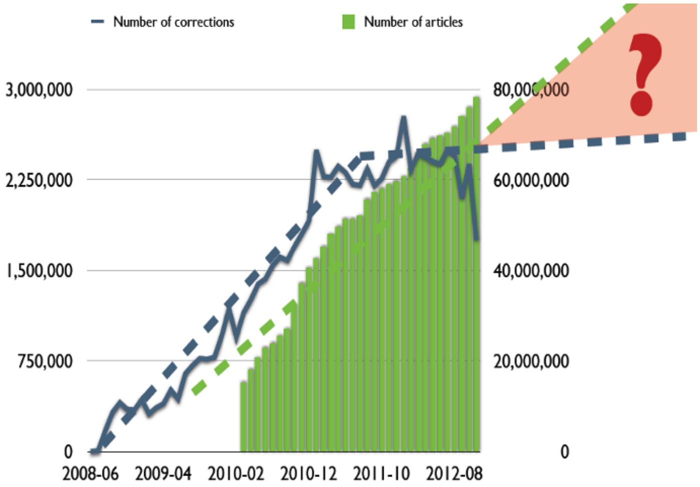


Figure 2.36. *Change in the amount of content compared to that of the number of corrections on TROVE, according to [HAG 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Contributors to the TROVE, Cambridge Public Library or California Digital Newspaper Collection projects are genealogists.

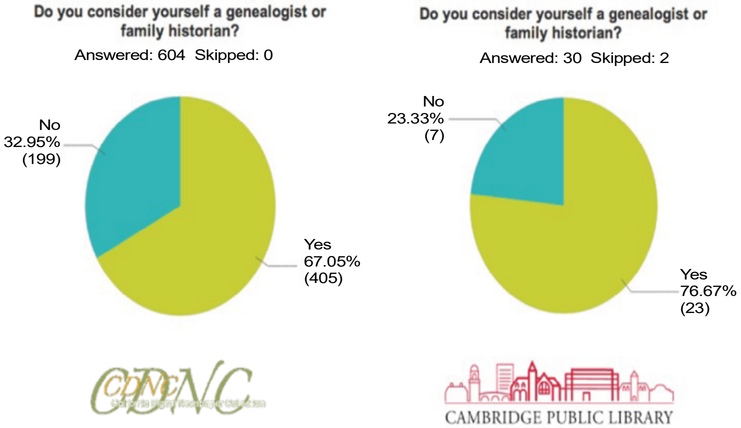


Figure 2.37. *Proportion of genealogists among the contributors, according to a CDNC/Cambridge Public Library survey. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

They are also more likely to be retirees (which is not the case with gamification) (see Figure 2.38).

In Figure 2.39, we can see that they are interested predominantly in genealogy.

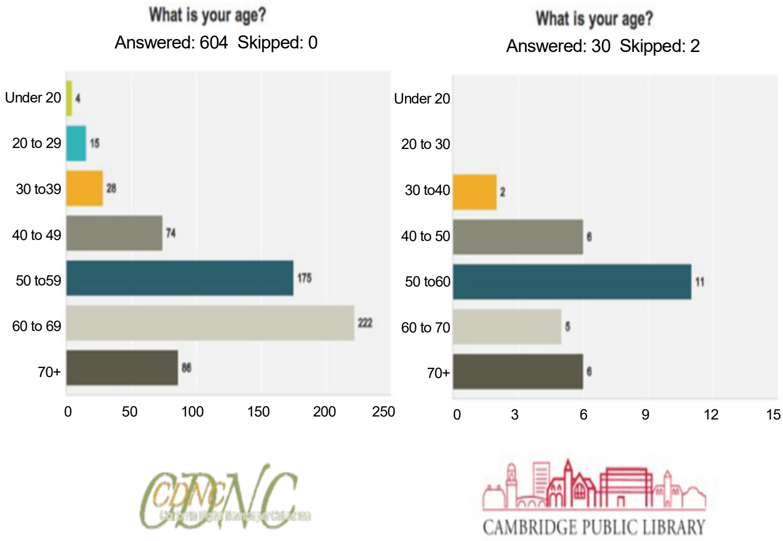


Figure 2.38. *Distribution of volunteers by age group, according to a CDNC/Cambridge Public Library survey. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

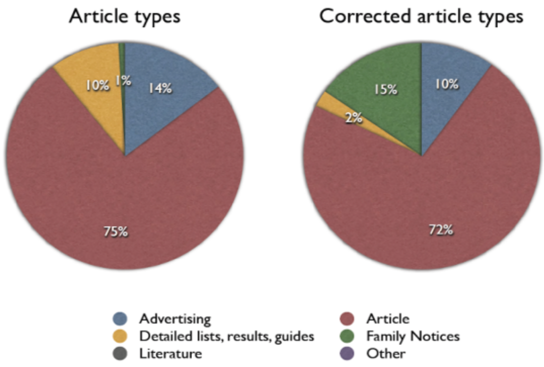


Figure 2.39. *The types of documents distributed on TROVE compared to the types of documents that are corrected there, according to [HAG 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

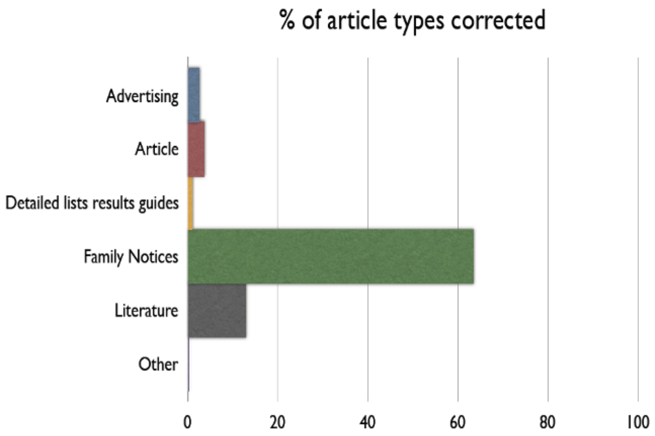


Figure 2.40. *Most corrected types of documents on TROVE, according to [HAG 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

If the largest part of volunteers is recruited from among well-off retirees interested in genealogy or local history, this raises the question of the continued existence of this type of service. When another generation of retirees has replaced the current one, nothing indicates that it will be equally passionate about local history. It also raises the question of representation of every part of the society that cultural institutions are supposed to serve.

Beyond the niche of crowdsourcing applied to libraries, this phenomenon that states that only a minority of Internet users are the source of the majority of the content also applies on the scale of Wikipedia where 90% of the contributions are the work of only 10% of users.

In the case of cultural institutions, in particular, it is therefore somewhat difficult to really talk about crowdsourcing, since it does not really involve a crowd of and anonymous, undifferentiated Internet users contributing irregularly or a limited, but rather small communities of loyal volunteers who help each other. The majority of successful crowdsourcing projects have not benefitted from large, anonymous crowds, but have managed to elicit the participation of a few engaged volunteers [OWE 13]. According to a study by [CAR 13], in over 36 crowdsourcing projects, the number of contributors were between a few hundreds and several tens of thousands, and the average number of participants around 5,000 or 6,000 Internet users.

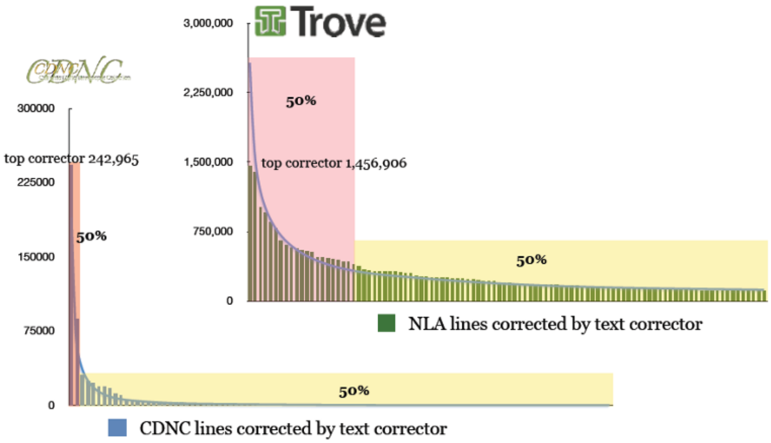


Figure 2.41. *Classification of contributors according to the number of lines corrected for the TROVE and CDNC projects, according to [ZAR 14]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Figure 2.42 shows a more original way to illustrate this phenomenon.

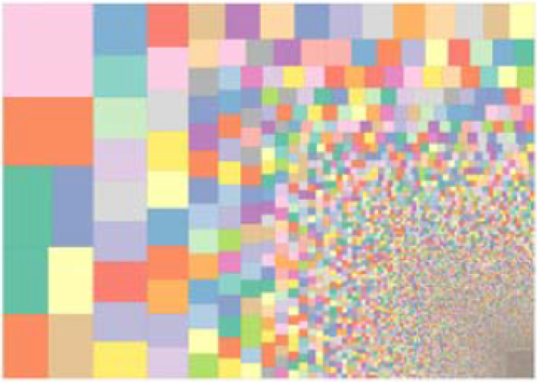


Figure 2.42. *Portion of the work accomplished by each contributor to the Old Weather project offering to transcribe meteorological observations, from Brumfield, manuscripttranscription.blogspot.fr, 2013. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

In the preceding diagram, each square corresponds to a volunteer and the size of each one is proportional to the quantity of transcriptions carried out.

This shows that out of the 1.6 million records produced by the British Royal Navy between 1905 and 1929 and transcribed by around 16,000 volunteers, 10% of the transcriptions are the work of only 20 very productive volunteers.

As a result, some authors such as [CAU 12a] prefer the term "*community sourcing*" to "crowdsourcing" to express the act of seeking to mobilize a community rather than a crowd of undistinguished Internet users.

2.5. Folksonomy, cataloguing and participative indexing

Folksonomy is a term invented in 2004 by Thomas Van der Wal from the words "*folk"* and "*taxonomy*". It is synonymous with *collaborative tagging*, *social tagging*, *social classification* and *social indexing*. The subject has already been widely studied and it would not have been appropriate to concentrate too much effort on the subject. We will consequently limit the subject of study to several of the representative and more original projects.

2.5.1. *Explicit crowdsourcing through volunteer tagging: Flickr: the Commons*

The site Flickr was created in 2004. It assembles a community of 51 million registered photographers. According to Yahoo, 4.5 million photographs are uploaded to the site each day [COL 13]. On January 16, 2008, the Library of Congress placed 3,000 photos on Flickr in order to increase its visibility and to allow indexing by Internet users, as well as comments, sharing, saving to favorites and reuse. The choice of Flickr rather than Picasa, Wikimedia or other sites was mostly justified due to the existence of a significant preexisting community. In 2013, 56 institutions in 14 countries, in addition to the Library of Congress, were also participants in the project.

Only 10 months after uploading the photos on Flickr, on October 23, 2008, the project had generated no less than 10.4 million photo views and around 6 million visits (equal to an average of 500,000 visits per month). Five years later, in January 2013, 250,000 images had been uploaded to Flickr. They had generated 2 million tags and 165,000 comments on the part of 165,000 contributors. Other institutions such as the Smithsonian stated that in only 3 months, their photographs had received an average of 2,348 visits per day, which is as many visits as in 5 years when the photos were previously on the institution's site. For the Smithsonian, between June and October 2008, 513 comments were added to added on 254 photographs this way (22% of the corpus) with an average of two comments per image and one comment per 2,089 visits.

Flickr was also used to collect photographs from Internet users by the Library and Archives Canada. "In more than a hundred years of existence, the various branches of the Library and Archives Canada (LAC) have collected more than twenty-five million photographs. It took only six years to see the website Flickr assemble five billion images" [CAS 11].

As with all of the crowdsourcing projects applied to cultural heritage, the main part of the content is produced by an active minority of Internet users. Regarding Flickr: The commons, 40% of tags are thus the work of only 10 *super taggers* who added more than 3,000 tags each [HOL 10a].

In this way, Internet users have allowed institutions participating in the project to identify people, places and events.

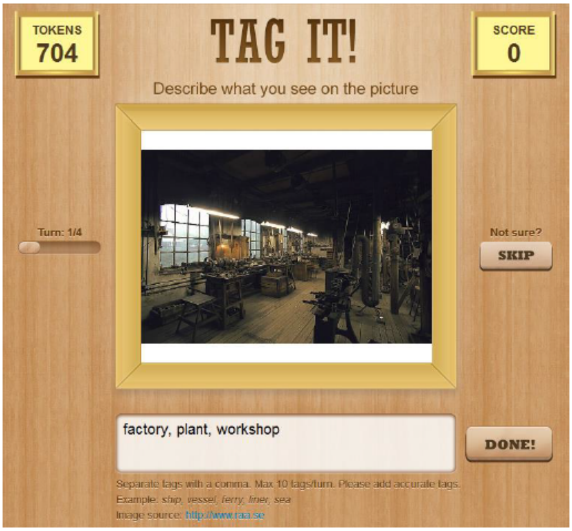
Regarding quality, a 2006 study by Guy & Tonkin and related by Earle estimates, using a sample, that only 40% of tags occurred in the *Open Source Aspell* dictionary [EAR 14].

2.5.2. *The use of gamification: Art Collector*

Art Collector is a game more specifically intended for digitized heritage and inspired by the experiments in gamification studied previously. It is a game developed on an experimental basis on Facebook around *Swedish Open Cultural Heritage* (SOCH), which agglomerates close to 100,000 images collected on various websites promoting Swedish cultural heritage.

As Paraschakis [PAR 13] notes, the choice of the social network Facebook comes from the very large amount of traffic generated by its games such as Mafia Wars or Farmville. The objective of the game is to build a collection of images and paintings and to become the biggest art collector by adding up the cumulative value of the works, with the value of a work being proportional to the number of tags which describe it.

There are two types of collections: private collections built by the players and public collections whose pieces still belong to no one. In order to acquire an image, the player must have offered more than half of its tags. At the end of this first round (*Tag It!*) in which four images are offered to each person (it is still possible to pass on an image), the players are rewarded with four tokens for each original tag and two tokens for each already-existing tag, i.e. one shared with one or several players. The goal of this first round is to offer as many keywords as possible. It is also possible to win 40 tokens if one of his or her social network friends agrees to participate. The three best players receive a medal (gold, silver or bronze). A player who has entered more than 50 tags obtains the *Power Tagger* medal and someone who as entered more than a hundred wins that of *Super Tagger*.



Describe what you see on the picture

factory, plant, workshop

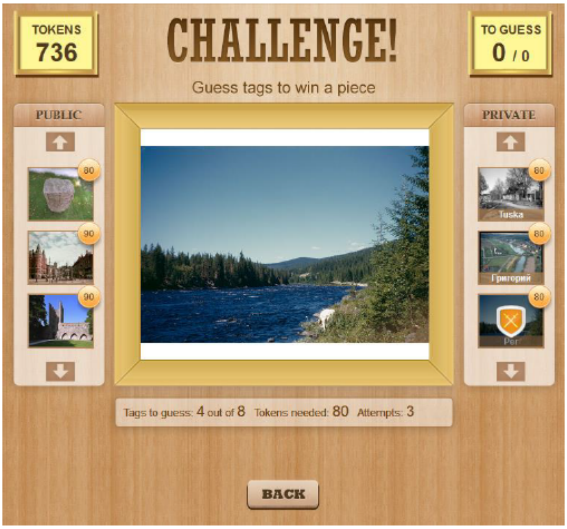
Separate tags with a comma. Max 10 tags/turn. Please add accurate tags.
Example: *ship, vessel, ferry, liner, sea*
Image source: http://www.raa.se

Figure 2.43. *Screen capture of the game Art Collector, first round, from [PAR 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

The works that receive at least four tags are then bought by the public collection for the second round (*challenge*).

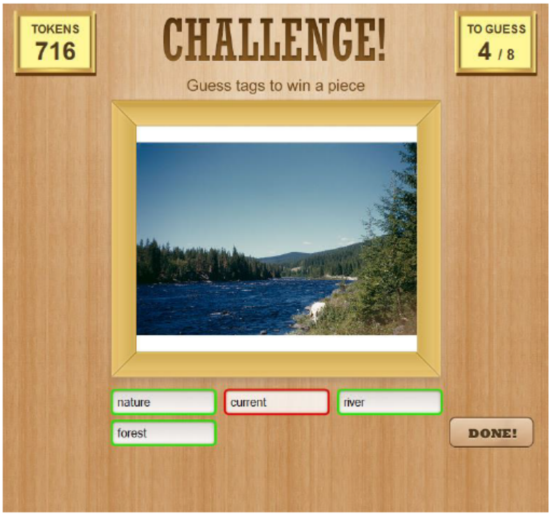
In this section, the players seek to win works that belong to the public collection or to private collections, especially, players who are part of the same social network, to enrich their private collections by guessing the keywords that correspond to them. Each attempt costs 20 tokens. The work is won if more than half of its tags have been guessed.

In the first part of the game (first round), indexings of the documents are produced. In the second round, these indexings are validated.



Tags to guess: 4 out of 8 Tokens needed: 80 Attempts: 3

Figure 2.44. *Screen capture of the game Art Collector, round 2, choice of a piece, from [PAR 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*



nature current river
forest

Figure 2.45. *Screen capture of the game Art Collector, round 2, trying to win a work, from [PAR 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

According to the statistics that have been published, 103 users had connected to the game in 2 weeks. Among these players, 56.3% of them played several times. Nevertheless, 35% of them did not add a tag. Sociologically, among the age groups of the players, the most numerous were the 25- to 34-year-old group. Men were 10% more numerous than women.

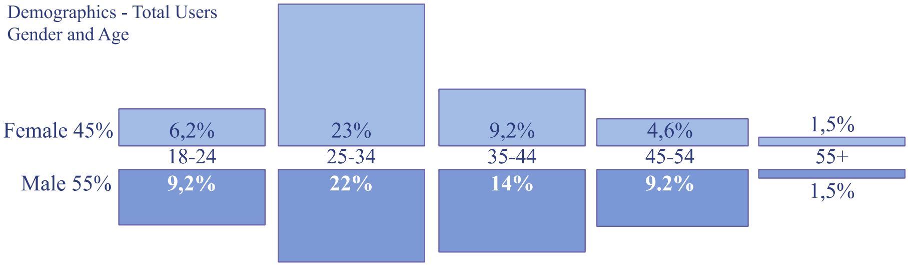


Figure 2.46. *Gender and age of the players of Art Collector, according to [PAR 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

As Paraschakis and Gustafsson Friberger [PAR 14] point out, the game Art Collector appeals to both the spirit of competition (table of the best players, results, challenges), communication (search for friends, notifications) and collaboration (sharing a trophy).

For a more complete overview, we could have also mentioned the 1,001 Stories Denmark project, the Netherlands Institute for Sound and Vision wiki, the British Library's Historical Maps Pilot, the Mtagger project at the University of Michigan, the PennTags project at the University of California, the Social OAC project at the Daniel Library, the Australian project Describe Me at the Victoria Museum, the Tag! You're It! and Freeze tag! Projects at the Brooklyn Museum, the British project Your Paintings Tagger and the British project Operation War Diary of the Imperial War Museums and the game Tagging Wasida with which Internet users gain points every time their tags are validated by others [SMI 11a].

The domain of folksonomy is too large and has been used for much too long for it to be useful and judicious, within this work this work, to seek to comprehensively study all of the of publications on the subject. That is why we have limited ourselves to the most representative initiatives and publications, significant and innovative with particular focus on gamification.

Tagging is a dialogue between the visitor and the work, between the visitor and the museum [TRA 06]. Although folksonomy offers total freedom to the user on which it is centered beyond any controlled, restrictive and expensive language, it can also be misused by those who are seeking to improve the referencing of a web page and become a source of infopollution and contribute to the creation of a "semantic Babel" [LED 06]. Thus, according to a 2006 study by Guy & Tonkin reported by Earle, only 40% of the tags occurred in the *Open Source Aspell* dictionary [EAR 14].

There are other projects that we could cite: the steve.museum, GLAM Wikimedia, Glashelder! and VeleHanden, museumgam.es, Metadata Games, SaveMyHeritage, Picaguess.

In order to avoid weighing down this second chapter relative to the overview of projects, we have only chosen the most significant projects in this chapter. Here is a list of other projects that we have identified:

– Addressing history, University of Edinburgh & National Library of Scotland (United Kingdom);

– Alto Editor IMPACT Centre of Competence;

– Civil War Diaries & Letters Transcription Project, The University of Iowa Libraries;

– Crowd4U (Japan): shared platform for crowdsourcing projects launched in 2010;

– Dickens Journals Online (United Kingdom): participative OCR correction;

– Family Search Indexing, Family Search (since 2004, 780,000 volunteers, 100,000 active volunteers per month, 1,500,088,741 records indexed in July 2012);

– FieldData, Atlas of Living Australia/Gaia Resources;

– Harold "Doc" Edgerton Project, MIT?;

– Islandora TEI Editor UPEI, Robertson Library;

– Itineranova-Editor, Stadsarchief Leuven/HKI Cologne;

– L-Crowd (Japan): launched in 2012 by Japanese university libraries for metadata correction;

– Metadata Games: a game developed by the Tiltfactor Laboratory (Dartmouth College) and offering Internet users the opportunity to *tag* photos, audio or video recordings of libraries or archives;

– Metadata Games project: a game developed by the University of Munich and offering a game involving *tagging* photos;

– Marine Lives (United Kingdom): transcription of English marine manuscripts from the 17th Century;

– National Archives Transcription Pilot Project, US National Archives;

– North American Bird Phenology Program, USGS;

– OpenScribe;

– Prism, University of Virginia (United States);

– Project Runeberg (see: runeberg.org, consulted June 23, 2016);

– PyBOSSA, Citizen Cyberscience Centre/OKFN;

– Scribe, Zooniverse;

– Scripto, Center for History and New Media at George Mason University;

– Son of Suda On-Line, Integrating Digital Papyrology;

– TextLab, John Bryant *et al.*, Hoftstra University;

– Unbindery, Ben Crowder;

– VdU-Editor Monasterium.net/HKI Cologne;

– Velehanden.nl;

– Veridian, DL Consulting;

– Virtual Transcription Laboratory, Poznań Supercomputing and Networking Center;

– Wiki::Score;

– Word Soup (see: cat.iti.upv.es/wordsoup, consulted June 23, 2016);

– World Archives Project, Ancestry.com.

In this overview of the projects most representative of crowdsourcing in digital libraries, we have addressed projects placing content online, digitization and printing on demand, OCR correction and then indexing. Regarding OCR correction in particular, we have differentiated between explicit crowdsourcing, gamification, implicit crowdsourcing and paid crowdsourcing.

In Chapter 3, which we will dedicate to a state the art of crowdsourcing in digital libraries, we will return to it by developing a taxonomy of projects. We will also address communication for recruiting, the types of motivations of contributors, their sociology, the quality of the contributions, their reintegration, their legal status, the evaluation of projects and change management.

3

Overview and Keys to Success

Based on an overview of projects and a reading of the literature on the subject, we have presented a summary of the subject in the form of an overview. This summary also contains original analyses that do not come from the literature.

3.1. Typologies and taxonomies of projects

Although we have already begun this work with a necessary introductory definition of crowdsourcing, we have been obliged, in this section dedicated to analyses in the area of library and information science, to revisit this definition. This time we will do so in a less general way and in a manner more relevant to the domain of digital libraries and by producing an original taxonomy.

Taxonomy, in natural science, consists of classifying species according to their traits and their characteristics into classes, orders, families and genera. This particular science has inspired other disciplines, especially library and information science. Within the domain of crowdsourcing, numerous taxonomies have been proposed in the literature, ones which we will summarize here before offering our own original classification of crowdsourcing projects in the field of digital libraries.

Initially, John Howe, inventor of the term crowdsourcing, distinguished the following four major types:

– *collective intelligence*: resolving problems using the wisdom of crowds;

Digital Libraries and Crowdsourcing, First Edition. Mathieu Andro.
© ISTE Ltd 2018. Published by ISTE Ltd and John Wiley & Sons, Inc.

– *crowdcreation*: using collective creativity;

– *crowdvoting*: asking the opinion and the advice of Internet users;

– *crowdfunding*: making use of participative financing.

In continuation of Howe, Harris returns to his taxonomy but more specifically distinguishes microtasks and macrotasks, the latter relying more on innovation via Internet users or via the wisdom of crowds of Internet users.

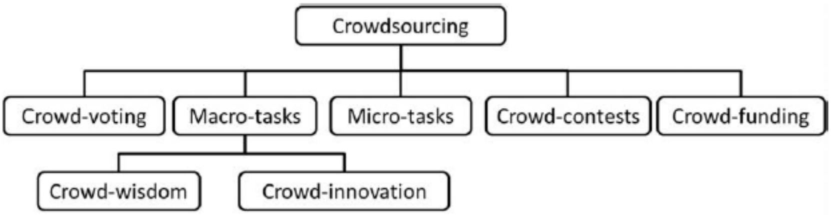


Figure 3.1. *Taxonomy of crowdsourcing, from [HAR 13]*

The majority of authors have also sought to classify projects according to levels of engagement and initiative, thus distinguishing engagement, participation, contribution and volunteering or, more simply, distinguishing participative or contributive crowdsourcing and collaborative crowdsourcing [BOE 12, BON 09, DUN 12, OOM 11, RAD 14, TWE 12]:

– participative crowdsourcing or contributive crowdsourcing: the public contributes simply by producing data within the framework of projects created and run by institutional investors. The public’s work is determined and limited and requires relatively low individual investment (microtasks);

– collaborative crowdsourcing or cocreation: the public takes a more active role in the decisions of the project’s collection management policy and the project calls for larger individual investment (macrotasks). Certain authors [BON 09, RAD 14] sometimes distinguish between collaborative crowdsourcing and cocreation:

- collaborative crowdsourcing: active partners interact with each other, but within a context controlled by the institution;

- cocreation: partners also participate in the policies and the definition of the project's goals or can even be the originators of projects.

More precisely, Bonney *et al.* [BON 09] present Table 3.1, which we have adapted to our field.

Scientific steps	Contributive projects	Collaborative projects	Cocreation projects
Choose and define a question	No	No	Yes
Gather information and resources	No	No	Yes
Make analyses and hypotheses	No	No	Yes
Conceive of data collection methods	No	Possibly	Yes
Produce data	Yes	Yes	Yes
Analyze data	Possibly	Yes	Yes
Interpret data and draw conclusions from it	No	Possibly	Yes
Broadcast the conclusions	Possibly	Possibly	Yes
Discuss the results and formulate new questions	No	No	Yes

Table 3.1. *Model of public participation inspired by [BON 09]*

Among the tasks identified in our overview (posting material online, digitization and printing on demand, OCR correction, transcription, indexing), obviously contributive projects are the only ones that currently exist in libraries. Only digitization on demand and posting content online could be considered collaborative to the extent that the public participates in building the collection and thus in the acquisition and collection management policies of the digital library.

According to [STI 14], five successive steps of engagement exist:

– individuals consume content;

– individuals interact with content;

– individual interactions are networked together;

– individual interactions are networked in social networks;

– individuals commit socially to each other.

It is also possible to refine the distinction between participative and collaborative according to the types of participation to the extent that, on the web, we already find the following categories according to the study "Forrester's NACTAS Q4 2006 Devices Access Online Survey" reported by [RAD 14]:

– the "creators" (13%) who publish websites or blogs and upload videos;

– the "critics" (19%) who comment and evaluate;

– the "collectors" (15%) who share on social networks;

– the "sociable" (19%) who use social networks;

– the "spectators" (33%) who read content on the Internet;

– the "inactive" (52%) who do not fit into any of the previous categories.

By drawing inspiration from this general classification and applying it to the cultural domain, we obtain the following categories of curators (minority), producers, commentators, sharers of content and consumers (majority) (from [RAD 14]).

One could also sort projects according to the following criteria:

– Who contributes? An indefinite and open crowd of Internet users (crowdsourcing) or a more specific and determined group, a community (community sourcing), the local populations?

– Why does the crowd contribute? What type of motivations do they have? Intrinsic motivations or rather extrinsic ones?

– How does the crowd contribute? Through competition or, conversely, through collaboration?

– For whom does the crowd contribute? For private interests or public interests?

– What is the project's main goal? Obtaining data or mobilizing the crowd around collections in order better raise awareness?

– What do the contributors provide? Money? (crowdfunding) Work? Knowledge? Ideas?

It is also possible to classify projects according to the level of interaction and competition of the crowd [REN 14b] by distinguishing:

– cumulative crowdsourcing: the juxtaposition and the aggregation of individual participations likely to lead to unexpected discoveries (“small streams become great rivers”[1]);

– collaborative crowdsourcing: orchestrated through the collaboration of individuals (“unity is strength”);

– competitive crowdsourcing: competition to be judged (“may the best man win”);

– “coopetitive” crowdsourcing: cooperation in a spirit of competition (“all for one, one against all”).

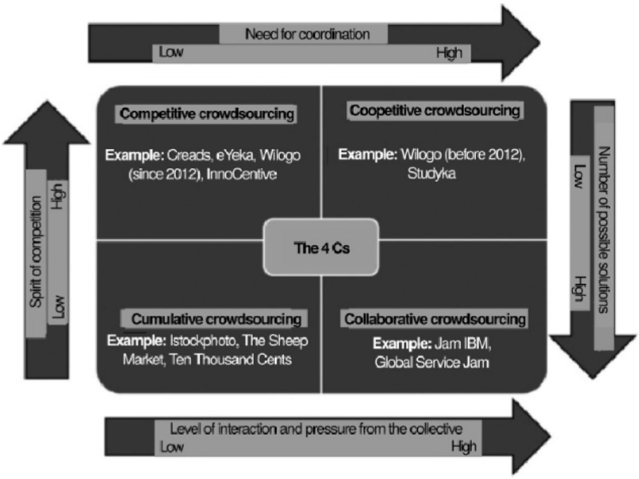

Need for coordination
Low
High
Spirit of competition
Low
High
Competitive crowdsourcing
Example: Creads, eYeka, Wilogo (since 2012), InnoCentive
Coopetitive crowdsourcing
Example: Wilogo (before 2012), Studyka
The 4 Cs
Cumulative crowdsourcing
Example: Istockphoto, The Sheep Market, Ten Thousand Cents
Collaborative crowdsourcing
Example: Jam IBM, Global Service Jam
Number of possible solutions
Low
High
Level of interaction and pressure from the collective
Low
High

Figure 3.2. *Taxonomy of the 4Cs of crowdsourcing, from [REN 14b]*

Finally, it is obviously also possible to classify projects according to the type of activity offered, as we have done elsewhere, in part, in our overview

[1] “You see few great rivers start at great sources; most are multiplied by the streams that flow into them”, Ovid, *Complete Works*.

of projects, distinguishing participative uploading, digitization on demand through crowdfunding, participative OCR correction and folksonomy.

In the case of the digitization of libraries, we can thus identify the following activities:

– selection of documents available to be digitized according to the legal criteria (the author has been dead for more than 70 years), relevance (document has not already been digitized elsewhere) or scientific and thematic criteria;

– material descriptions of the documents to be digitized (format, number of pages, angle of opening, condition of the document);

– digitization (organization of streams of digitization, digitization, dispatching the streams);

– production of raw OCR (with optical character recognition software);

– quality control of the digitization (monotonous work usually handled internally by libraries which control certain points in all of deliveries or the samples. This work can sometimes be subcontracted by a service provider that controls the work of the original service provider);

– uploading documents online;

– cataloging or recataloging of documents put online;

– indexing or reindexing of documents;

– OCR correction (generally done for editorial projects, the production of EPUB or MOBI files or text mining projects with the help of the service provider using low-cost labor in Madagascar or India, for example);

– perennial archiving (transfer of high-resolution files in conservation formats to archiving servers accompanied by metadata and bibliographic techniques);

– editorial development and contextualization by adding a scholarly apparatus for each text (bibliographic information, summaries, table of contents, news, analyses, associated documents, etc.);

– creation of electronic books readable on tablets in EPUB and/or MOBI formats.

We could have also added the following activities: creating links, commenting, categorizing, cataloguing, contextualizing, georeferencing and translating by using the classification proposed by Dunn and Hedges [DUN 12]. Other classifications, according to the type of documents (images, texts, manuscripts, videos, sounds, maps) or according to the type of data produced by contributors (texts translated or transcribed or corrected, metadata, summaries, knowledge, money, etc.), would also be conceivable.

As part of our overview of projects, we have also been led to distinguish in a more original way, the types of crowdsourcing themselves:

– explicit crowdsourcing:

- free explicit crowdsourcing (use of volunteer Internet users);

- paid explicit crowdsourcing (use of paid Internet users);

– implicit crowdsourcing (use of the involuntary work of Internet users);

- gamification (use of the work of Internet users in the form of games);

- crowdfunding (use of the financial contributions of Internet users).

This taxonomy is original. The distinction between implicit and explicit crowdsourcing is drawn from [HAR 13] who, regarding volunteer participation, talks about explicit crowdsourcing and, concerning involuntary participation, above implicit crowdsourcing.

From this taxonomy comes an analysis of the literature and the taxonomies previously produced; we have tried to cross these various forms of models with the different activities of a digital library development project. By crossing them, we have sought to identify possible forms of crowdsourcing that remain to be invented and which have not, to our knowledge, been the subject of experiments.

On the vertical axis of Table 3.2, we find the different tasks of a digitization project:

– the selection of documents that deserve to be digitized according to scientific and historic criteria and after verification that they have not been already and that they can be from a legal perspective;

– digitization;

– financing;

– quality control of the digitization, OCR, metadata;

– cataloging;

– indexing;

– OCR correction of print materials;

– transcription of manuscripts.

On the horizontal axis, we find our taxonomy:

– explicit crowdsourcing;

– gamification;

– implicit crowdsourcing.

And, for each category, we have created subcategories to distinguish unpaid and volunteer work from paid work, although there exist a multitude of intermediate forms. We have also chosen to distinguish the quantitative degrees of engagement:

– participative: Internet users produce data in the form of microtasks and with relatively low commitment to the institutions within a limited framework;

– collaborative: Internet users participate in the policies and in defining the goals of the project, and commit more strongly.

Thus, if we cross the types of activities linked to digitization projects with all of these variables, we obtain the following original taxonomic matrix. This matrix is in the interest of allowing the identification of new forms of crowdsourcing applied to digitization projects, which remain to be invented. Its limit rests in the artificial and sometimes meaningless nature of certain combinations. Thus, all of the forms in red (see color section) do not seem to us to be able to find an application and they remain in the majority in Table 3.2.

Among all of this systematicity of the forms of crowdsourcing likely to exist in the field of digitization in libraries, some, in red (see color section), do not make, from our point of view, unfortunately very much sense and would probably not have an application in the future.

Other forms of crowdsourcing already exist as we can see in Table 3.3.

	Explicit crowdsourcing				Gamification				Implicit crowdsourcing			
	free		paid		free		paid		free		paid	
	participative	collaborative	participative	collaborative	participative	collaborative	participative	collaborative	participative	collaborative	participative	collaborative
Selection	111a	112a	121a	122a	211a	212a	221a	222a	311a	312a	321a	322a
Digitization	111b	112b	121b	122b	211b	212b	221b	222b	311b	312b	321b	322b
Financing	111c	112c	121c	122c	211c	212c	221c	222c	311c	312c	321c	322c
Quality control	111d	112d	121d	122d	211d	212d	221d	222d	311d	312d	321d	322d
Cataloging	111e	112e	121e	122e	211e	212e	221e	222e	311e	312e	321e	322e
Indexing	111f	112f	121f	122f	211f	212f	221f	222f	311f	312f	321f	322f
OCR correction	111g	112g	121g	122g	211g	212g	221g	222g	311g	312g	321g	322g
Transcription	111h	112h	121h	122h	211h	212h	221h	222h	311h	312h	321h	322h

Table 3.2. *Activities of a digitization project crossed with the types of crowdsourcing. For a color version of the table, see www.iste.co.uk/andro/libraries.zip*

Code	Type	Form
111a	Free and participative explicit crowdsourcing applied to the selection of documents	Suggesting the digitization of a document
311a	Free and participative implicit crowdsourcing applied to the selection of documents	Using consultation and borrowing statistics from library catalogs to identify the documents to be digitized
111b	Free participative explicit crowdsourcing applied to digitization	Using amateur digitizations of books or archives by Internet users to enrich digital libraries (Internet Archive, for example)
111c	Crowdfunding applied to digitization	Digitization of documents on demand using participative financing (Numalire, for example)
111f	Free participative explicit crowdsourcing applied to indexing	Tagging (folksonomy, Steve Museum, for example)
211f	Free participative gamification applied to indexing	Games involving the indexing of digitized documents (Google Image Labeler, for example)
111g	Free and participative explicit crowdsourcing applied to OCR correction	Participative OCR correction (Wikisource, for example)
121g	Paid and participative explicit crowdsourcing applied to OCR correction	Paid OCR correction by Internet users (on the Amazon Mechanical Turk Marketplace, for example)

211g	Free participative gamification applied to the OCR correction	Gamification around OCR correction (Digitalkoot, for example)
311g	Free and participative implicit crowdsourcing applied to OCR correction	The use of entries from Internet users for security reasons (reCAPTCHA, for example)
111h	Free and participative explicit crowdsourcing applied to manuscript transcription	Participative transcription of manuscripts (Transcribe Bentham, for example)
121h	Paid and participative explicit crowdsourcing applied to manuscript transcription	Paid transcription of manuscripts (on the Amazon Mechanical Turk Marketplace, for example)

Table 3.3. *Existing types of crowdsourcing applied to digitization*

Others still have not yet been identified or still remain to be invented (see Table 3.4).

Code	Type	Form
112a	Free and collaborative explicit crowdsourcing applied to the selection of documents	Allowing Internet users to influence the digitization policy directly
121a	Paid and participative explicit crowdsourcing applied to the selection of documents	Paying Internet users to identify the documents which deserve to be digitized by finding the death dates of the authors and by verifying whether or not they have been already digitized
211a	Free and participative gamification applied to the selection of documents	Creating a game in which Internet users must find out if the document can be digitized by adding note on its interest, finding the death dates of the authors or by verifying that it has not already been digitized
221a	Paid and participative gamification applied to the selection of documents	Paying the best players of the game mentioned previously
121b	Paid and participative explicit crowdsourcing applied to digitization	Paying Internet users or readers for the previously unpublished documents that they digitize and post online on a digital library
311b	Free and participative implicit crowdsourcing applied to digitization	Automatically conserving images of the photocopies made by Internet users systematically associated with document references identified via RFID in order to be able to then place them in digital libraries

111d	Free and participative explicit crowdsourcing applied to quality control of the digitization	Asking Internet users to validate the quality of a particular digitized page as a particular control point and compare their validations
121d	Paid and participative explicit crowdsourcing applied to quality control of the digitization	Paying Internet users for this validation work
211d	Free participative gamification applied to quality control of the digitization	Creating a game in which Internet users validate the quality of the digitized pages based on particular criteria
221d	Paid and participative gamification applied to quality control of the digitization	Compensating the best players of the game mentioned previously
111e	Free and participative explicit crowdsourcing applied to the cataloguing of digitized documents	Asking Internet users to catalog digitized documents
121e	Paid and participative explicit crowdsourcing applied to the cataloging of digitized documents	Compensating Internet users for this cataloging work
211e	Free participative gamification applied to the cataloging of the digitized documents	Making a game out of catalog digitized documents
221e	Paid participative gamification applied to the cataloguing of digitized documents	Paying the best players of the previously mentioned game
121f	Paid and participative explicit crowdsourcing applied to indexing	Paying Internet users for their keywords and their tags
221f	Paid participative gamification applied to indexing	Paying the best players of tagging games
221g	Paid participative gamification applied to OCR correction	Paying the best players of the OCR correction games
211h	Free participative gamification applied to the transcription of manuscripts	Making a manuscript transcription game modeled on those that already exist for OCR correction
221h	Paid participative gamification applied to the transcription of manuscripts	Paying the best players of manuscript transcription games
311h	Free and participative implicit crowdsourcing applied to the transcription of manuscripts	Use the reCAPTCHA system for manuscripts

Table 3.4. *Types of crowdsourcing applied to digitization that remain to be invented*

To summarize the above, here is the original taxonomy that we offer in Table 3.5.

Types of crowdsourcing		Definition	Examples
Explicit crowdsourcing	Free	Use of voluntary work by volunteer Internet users	TROVE
	Paid	Use of voluntary work by paid Internet users	Amazon Mechanical Turk Marketplace
Implicit crowdsourcing		Use of the involuntary work of Internet users	reCAPTCHA
Gamification *human computation* *games with a purpose*		Use of the work of Internet users in the form of games	Digitalkoot
Crowdfunding		Use of the financial contributions of Internet users	eBooks on Demand

Table 3.5. *Taxonomy of crowdsourcing applied to digitization*

3.1.1. *Explicit crowdsourcing*

3.1.1.1. *Free explicit crowdsourcing*

This form of crowdsourcing is the oldest and the most widespread. It consists of using the free volunteer work of Internet users to add digitized documents to a digital library, correct OCR or transcribe writing, add metadata and add keywords.

3.1.1.2. *Paid explicit crowdsourcing*

This form of crowdsourcing, still not very widespread in libraries, consists of asking Internet users to do the same kind of work, but be paid for it. The rare experiments related in the literature and that we have mentioned in our overview, have been carried out on Amazon Mechanical Turk Marketplace or CrowdFlower.

3.1.2. *Implicit crowdsourcing*

This form of crowdsourcing that is even less widespread in libraries is not, to our knowledge, used outside of the reCAPTCHA project that makes it possible to have Internet users involuntarily correct the OCR of 30 million books digitized by Google Books when they enter distorted words to prove that they are not robots when they create accounts.

3.1.3. *Gamification*

This form of crowdsourcing consists of asking Internet users to produce work while playing. As we saw in the overview of projects, there are multiple experiments applying gamification to the digitization of libraries.

If we consider that many tasks still remain impossible for computers to do while not impossible for humans, and that the latter spend an increasing amount of their time playing in front of a computer, it is clear that there is an opportunity to use human intelligence for all kinds of expensive tasks and mobilize it in the form of games.

According to [PAR 13], online gaming is located just behind social networks among the most common activities on the web. Games such as Yahoo! Games, MSN's The Zone or Pogo.com frequently gather more than 100,000 visitors. According to [PAR 13], games on social networks attract 120 million people among which there are 81 million who play every day and 49 million several times per day. The Facebook game Farmville, in particular, attracts 83 million players per month and the Mafia Wars game, for its part, attracts 25 million per month. Facebook remains the leader with 91% of players, ahead of Google+ (17%), MySpace (15%) and Bebo (7%). According to [VON 08] citing a report from the Entertainment Software Association, 200 million cumulative hours are spent each day on video games in the United States, 65% of American households play video games and a United States citizen has already spent, on average, no less than 10,000 h playing video games by the time he has reached the age of 21. These 10,000 represent the equivalent of 5 years of full-time work, i.e. 40 h of work per week. More than half a billion people play, throughout the world, games on the web and do this for at least 1 h every day. In the United States, there are 183 million of them [EIC 12]. Regarding casual games such as puzzles, Solitaire, Patience or Minesweeper, there are 200 million people around the world who play them [RID 11]. A 2006 study by the company PopCap, reported by this author, revealed that 76% of players were women whose average age was 48 years old.

In the context the gamification of culture, where pleasure seems to have a growing importance in society, studies, like work, could be seen as a succession of challenges, with tests, quests, changing levels, points and bonuses. Organizations could therefore take their inspiration from video games to increase the motivation of their students or their collaborators. It would thus be possible to reuse the main resources and mechanisms of video games to reuse it in other contexts. Gaming is a voluntary, autonomous activity that makes it possible to have new experiences, to test and gain skills in a safe environment and without possible negative consequences. According to [CHR 11], gamification makes it possible to obtain better results than traditional crowdsourcing in terms of participation. Indeed, individuals are generally hesitant to dedicate a large part of their time to accomplishing difficult work or completing thankless tasks. However, they also sometimes have difficulty stopping playing video games. It can therefore be a good idea to transform thankless tasks into video games. This process, which consists of converting productive activities into games, is known as gamification. Gamification could also be defined as the act of applying elements of design, psychology and video game mechanisms in other contexts [DET 11b].

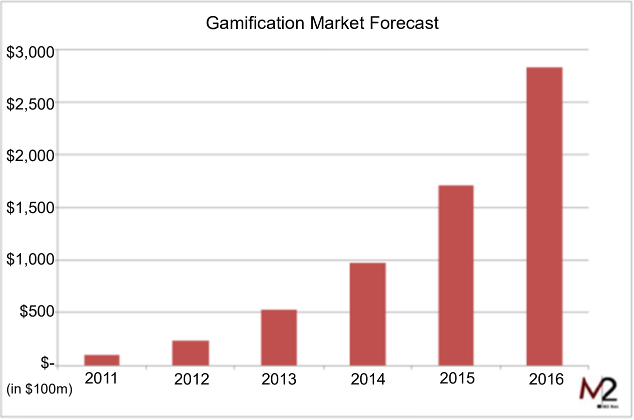
The term “gamification” was proposed by Nick Pelling in 2002, while the term human computation was proposed by Luis Von Ahn in his 2005 thesis. The term “games with a purpose” (GWAP) was proposed in 2008 by Von Ahn and Dabbish. As [VON 06a] suggests, it is enough to consider human brains as many processors in a network within a distributed system. Thanks to this system, each individual could participate in producing a massive calculation. Quinn and Bederson [QUI 11] produced a specific contribution in order to define the concept of human computation by taking up these different definitions. Drawing upon his works, we could define human computation as the use of mobilized collective human intelligence, by games, in order to resolve problems that computers do not yet have the capacity to take care of, which cannot be resolved by such limited groups of humans. Just as crowdsourcing replaces salaried employees with Internet users, human computation replaces computers with humans.

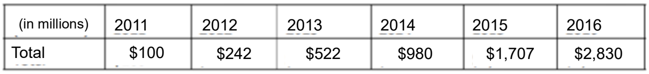
Gamification’s potential is very large. Von Ahn and Dabbish [VON 04] claim that the entirety of the images on Google Images could be indexed in 31 days by 5,000 Internet users playing the ESP Game for 24 h, seven days a week. He also reported that 1,000 players could index 12,000 images per day if each one dedicated 1 h of the day while it would take a traditional employee tagging 900 images per day over more than 125 days, close to four months, of full-time work to obtain the same result at the risk of causing a burn-out. In the field of participative sciences, if the hundreds of millions of people playing video games throughout the world who spend 3 billion hours per week playing spent only 1% of this time on the game fold.it, the significant results obtained in three years of the project could be obtained each week [GOO 11].

More recently and from a wider point of view, according to a press release from the company Gartner published in 2011, more than 50% of organizations managing the innovation process could incorporate gamification-related mechanisms in their businesses from then until 2015. We nevertheless see that in the same way as gamification is regularly confused with serious games, which only aim to train individually through the game and not to produce data, it is also regularly confused with “pointification”. Yet, assigning points for all kinds of actions has nothing to do with gamification. Games require a duration and a space; they are governed by rules, are subject to a purpose and utilize volunteers.

Gamification has already found numerous applications in multiple domains such as the indexing of videos or images, translation, transcription, summarizing documents, teaching and even video surveillance.

Its potential in the domain of digitization in libraries could be all the more important than the score of the participants being displayed. In the field of cultural institutions, we could mention DigiTalkoot (National Library of Finland) for OCR correction, Alum Tag (Rauner Special Collections Library, Dartmouth College) for indexing photographs, “Tag! You’re it!” (Brooklyn Museum) for indexing objects or Waisda? (Netherlands Institute for Sound and Vision) for the annotation of television broadcasts.





(in millions)	2011	2012	2013	2014	2015	2016
Total	$100	$242	$522	$980	$1,707	$2,830

Figure 3.3. *Time evolution since 2011 and forecast of the future gamification market, from [OLL 13]*

Unlike traditional explicit crowdsourcing, which appeals to altruistic feelings, gamification instead appeals to Internet users' desire to have fun. Harris [HAR 13] considers gamification as being at the intersection between serious games and crowdsourcing. Like serious games, gamification can also be very "serious"; the data produced while having fun could have a very serious use and be used by very serious organizations. However, gamification is different from serious games, since its purpose is utilitarian for the user who expects to gain individual benefit in terms of personal development, knowledge and training while with gamification, he seeks mostly to have fun while achieving a goal external to himself. In this respect, the Internet user's goal is not, unlike in a serious game, to receive training through a game, it is rather to produce useful data while having fun. Finally, gamification, unlike serious games, works based on microtasks autonomously compared to the each other and does not offer a scenario built in a linear way as is generally the case with serious games.

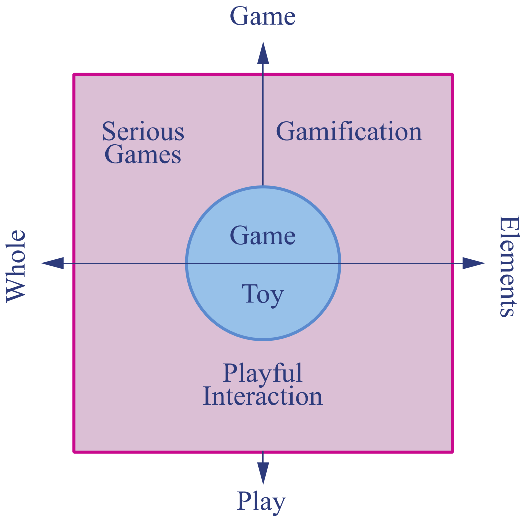


Figure 3.4. *Serious games and gamification, from [DET 11a]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

According to [VON 08], three major types of gamification exist:

– *output-agreement games*: each player has the same input information (for example, the same image for ESP Game) and the results produced can be different;

– *input-agreement games*: these games compare the entries of players such as in TagATune (a game which sends music to two Internet users who must communicate in writing to figure out whether or not it is the same clip);

– *inversion-problem games*: the first player has access to the whole problem and tries to make the second player find the solution such as in Peekaboom.

Following this, Quinn and Bederson [QUI 11] have sought to offer a classification of human computation projects according to the following broad characteristics:

– the type of motivations of Internet users:

- money or gratification;

- altruism;

- entertainment (gamification);

- reputation;

- implicit work (with reCAPTCHA, for example, the Internet user does not know that he or she is working for Google Books);

– the type of quality control:

- output agreement between the contributions of several Internet users who work independently and simultaneously;

- input agreement;

- financial incentives;

- design of the tasks in such a way that it is not easier to cheat than to actually perform the task;

- reputation with evaluation of Internet users' contributions the way eBay sellers are;

- redundancy (identifying the bad contributions and bad contributors via a voting system);

- traps (inserting intentional errors to verify that the work has been carried out thoroughly);

- statistical filtering;

- examination at several levels (a group of Internet users verifies the work of the first group. In digitization markets, for example, it happens that a service provider performs quality control for an original service provider);

- control by an expert;

- automatic control by programs and algorithms;

– the type of human skills mobilized:

- visual recognition;

- language comprehension;

- basic human communication;

– the order of the process:

- computer then worker then sponsor;

- worker then sponsor then computer;

- computer then worker then sponsor then computer;

- sponsor then worker;

– the architecture of requests for tasks:

- one to one;

- several to several;

- several to one;

- few to one.

Here are the recurring functions in gamification projects according to [GOT 14, HAM 14]:

– social functions:

- possibility of sharing on a social network, adding a like and in that way, making relatives and others aware of the existence of the game and, in particular, the display of Internet users appearing in the player’s social network and making it possible to offer to play with them all the while promoting the game virally;

- possibility of chatting and sending messages;

- micropayments allowing Internet users to financially support the development of games;

– functions of the games:

- statistics, number of points, medals, grades, rankings, rewards, challenges, competitions, challenges, goals;

- possibility of playing with other players without simultaneity, by simulating real time;

- time limit, stopwatch.

In terms of results, traditional crowdsourcing and gamification have been compared in several studies. McCarthy [MCC 12] sought to compare empirically the results between traditional participative OCR correction with the results obtained with gamification based on the Digitalkoot model in order to verify if gamification could have the effect of increasing the motivation of participants. Over the course of the experiment, two groups were then created. With gamification, we obtain 20% more participation according to his conclusions. In the same way, according to [FLA 12], games make it possible to collect more keywords per person. Thus, we obtain an average of six tags per visitor for the Library of Congress Flickr project versus 84 tags per visitor for the Tiltfactor Metadata Game.

According to a study conducted by [SAB 13] from an analysis and a summary of the literature on the subject, traditional crowdsourcing would be much less expensive and would require less time to be set up compared to games. According to this author, the motivation of volunteers would be easier to maintain. Finally, traditional crowdsourcing would be better perceived, from an ethical perspective, by the public and would benefit from a better image. In comparing the results obtained via a game and via the Amazon Mechanical Turk Marketplace, the authors of the study estimate that the game makes it possible to mobilize a less diversified variety of contributor profiles than with paid crowdsourcing via the Amazon Mechanical Turk Marketplace. Nevertheless, the game makes it possible to subcontract much more complex tasks and to obtain higher production quality; it would also have a slightly lower cost per task and it is less conducive to fraud.

With the game, players are more motivated by intrinsic reasons (amusement) while on the Amazon marketplace extrinsic motivations (financial reward) dominate. It could therefore be a good time to experiment with a game that appeals to two types of motivations with a game that offers a financial reward to the best players. They are thus simultaneously motivated by reasons as much intrinsic as extrinsic.

According to [HAR 13], resorting to gamification rather than traditional crowdsourcing makes it possible to improve the speed and the quality of the contributions, but would be more expensive and take longer to set up. Göttl [GOT 14] also estimates that games with a purpose (GWAP) are especially expensive to develop. As part of the thesis [HAR 13], the author sought to compare the results obtained for the identification of acronyms according to these two methods of contribution and between students and workers in the Amazon Mechanical Turk Marketplace. For gamification, the players were timed, evaluated in real time and ranked at the end of the section. He noted a greater precision in the identification of acronyms. However, students stand out as having a stronger ability to resolve the most difficult identifications. According to this student, gamification should therefore be favored for the simplest, most tedious tasks, and those which do not require too much concentration.

3.2. Communication and marketing for recruiting volunteers

If resorting to Internet users makes it possible to benefit from a form of voluntary and free work, institutions cannot ignore that significant expenses need to be authorized in order to develop platforms to recruit volunteers. Significant investments will therefore be necessary. Nevertheless, cultural institutions already enjoy a positive public image with the public. As public services, they appear to be trustworthy, without commercial motivations, and to be working in the service of the public interest. They often already have extensive experience in mobilizing volunteers, and planning events. Among the means of communication used by cultural institutions, we can list:

– putting up stickers, displays, the production of posters;

– academic articles;

– flyers distributed at trade fairs;

– conferences and conventions, organization of public meetings or of events and identifying and contacting people likely to contribute [BAU 10];

– using town councils, schools and organizations, the mobilization of preexisting communities;

– production of videos, widgets and teasers in order to increase the site's web traffic;

– use of mailing, active presence active on social networks (Twitter, Facebook, Vimeo, LinkedIn);

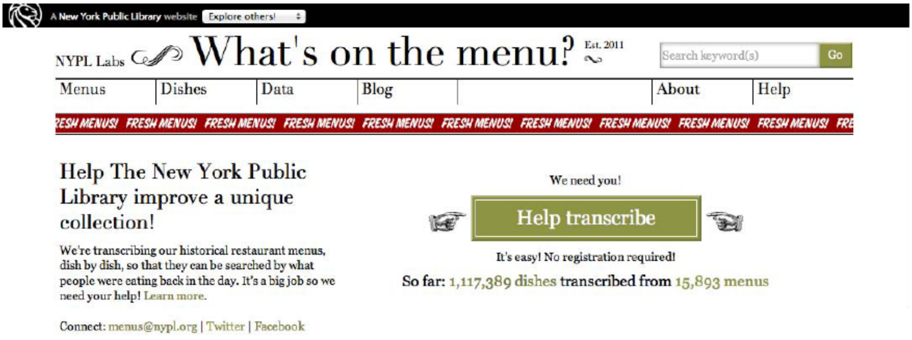
– use of web traffic already generated by the institutional site and its online catalogue;

– Transcribe Bentham even experimented, unfortunately without great success, with buying words as part of a Google Adwords campaign;

– more conventionally, the traditional media was also used successfully (press campaigns with announcements in the specialized, local, national press, community newsletters, radio and television broadcasts).

Donelle McKinley is a doctoral student at Victoria University who works specifically on interface design of crowdsourcing sites. According to her recommendations [MCK 13], a crowdsourcing site must have a home page that describes the project and invites volunteers to participate, and other pages to instruct volunteers in the execution of tasks. Internet users need to have a clear idea of why they are there and what they have to do. In order to convince an individual to collaborate, he or she must be interested in the subject, have the impression that his or her participation will be useful, that the project is feasible, that he or she will be supported, will be able to obtain responses to his or her questions, will have access to assistance, support forums, mailing lists, and will be recognized for his or her work. Still other pages are dedicated to registering volunteers. They must there present detailed information about the project, its team, its development status, and provide access to the profiles of other volunteers. Finally, Donelle McKinley recommends minimizing the user's effort, to allow rapid integration of new contributors without prior training because of an intuitive and ergonomic system. Thus, some Internet users might have only a few minutes to spend on a project, but it is necessary to be able to capture these precious minutes, considering that these Internet users can be legion. To be able to contribute, it would not be necessary to have made more than three clicks, as McKinley [MCK 12b] points out.

The content of the communication must be simple, clear, short and volontarist. Thus, as McKinley [MCK 12c] points out as part of the *What's on the Menu?* project the phrase "Help the New York Public Library to enrich a unique collection" is simultaneously short and simple, but it makes it possible to both say what the project is, who the sponsor is, who to address, how to participate, what the goal is and that it is the reason to participate. Expressions such as these could also be used: "Help us to create open and free access to the printed cultural heritage of the Library", "thanks to the efforts of people like you", "volunteers from around the world", "XX% of the collection has been corrected thanks to you. There is no more than XX% left to correct", etc. The Cleveland Museum of Arts invites Internet users to add keywords to the works that it broadcasts on the web by displaying the message "help others find this work" [TRA 06].


A New York Public Library website
Explore others!
NYPL Labs
What's on the menu?
Est. 2011
Search keyword(s)
Go
Menus
Dishes
Data
Blog
About
Help
FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS! FRESH MENUS!
Help transcribe

Help The New York Public Library improve a unique collection!

We're transcribing our historical restaurant menus, dish by dish, so that they can be searched by what people were eating back in the day. It's a big job so we need your help! Learn more.

Connect: menus@nypl.org | Twitter | Facebook

We need you!

It's easy! No registration required!

So far: 1,117,389 dishes transcribed from 15,893 menus

Figure 3.5. *Screen capture of the What's on the menu? press release: "Help the New York Public Library improve a unique collection. We need you! Help transcribe. It's easy! No registration required!" from [VER 13]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

As part of the Steve Museum project, the Cleveland Museum of Art invited its visitors to add keywords by employing this slogan: "Help others find this object" [CHU 06].

To measure the significance of the words chosen to recruit volunteers, Jeff Howe related a project in which Internet users were invited to become citizen journalists. However, unfortunately, no one clicked on the phrase "Be a Citizen Journalist". When the phrase was changed to "Tell Us Your Story" in order to invite Internet users to tell a story, the result was not any better. On the other hand, when the phrase was replaced by "Get published", Internet users finally flocked to the project in droves [ORG 10].

In order to recruit contributors, certain institutions can also, because of a monitoring system, monitor social networks, especially Twitter, Facebook, mailing lists and forums, where we talk about them or their collections and thus identify potential contributors to recruit.

From a marketing perspective, several social psychology techniques can be mobilized to increase the number of participations:

– the labeling technique that consists of asserting that we already think that the Internet users to whom we address as benefactors who are already positively open to the appreciation of cultural heritage;

– the "foot in the door" technique that consists of offering tasks that are very easy to complete in order to inspire an initial symbolic act of commitment, which will generally be followed by a larger commitment. The simple act of registering on a site is moreover already an act of involvement;

– the "foot-in-mouth" technique that consists of gaining the involvement of Internet users by politely asking them for personal information at the beginning of the interaction;

– the "but you are free to" technique that consists of reminding the Internet user that he or she is free to accept or refuse to participate;

– the "a little bit is better than nothing" technique. By claiming, for example, that even 10 min of the Internet user's time would still help the library considerably.

3.3. The question of motivations

If, unlike salaried employees, volunteers do not necessarily expect a financial reward in return for their contributions, they must nevertheless benefit from a return on the part of the institutions that benefit from their work. Crowdsourcing projects are always carried out necessarily for the mutual benefit of the institution and the Internet user.

The question of the motivations of contributors to crowdsourcing projects is recurring in the literature. We generally distinguish between intrinsic motivations and extrinsic motivations, which can moreover predominate in a very different way from one individual to another.

Intrinsic motivations, within the individual, are those that motivate him or her to act only out of interest in the work and pleasure that it provides. The activity is thus an end in itself, art practiced for art's sake, only for its content and only for the satisfaction drawn from it, for the beauty of the gesture, for self-actualization or the responsibilities, in a passionate and altruistic way, without seeking recognition or reward which risks, on the contrary, a decrease in motivation. The activity is performed for pleasure, for curiosity, for a feeling of competence, a search for a purpose, a feeling of freedom and self-determination.

On the contrary, extrinsic motivations, outside of the individual, are those that push him or her to perform an activity in order to obtain a result outside of this activity, to seek the effects and consequences of the activity outside of the activity itself such as recognition, reward or payment. It is therefore more restrictive and less free. The activity is thus an instrument, a simple means to achieve an end and obtain the desired result (in the realms of work, money or avoiding punishment).

The motives likely to encourage Internet users depend on the diversity of psychologies, cultures and social classes of the individuals. It is therefore suitable to take into account the diversity of profiles and the diversity of motivations likely to motivate them [SMI 13].

Drawing upon the different taxonomies and studies of the motivations of contributors found in the literature [ALA 12, ALA 13b, DUN 13, DWO 12, KAU 11, OWE 13, ROU 10, SMI 13], here is an original taxonomy that we offer in Figure 3.6.

These major types of motivations will be developed in sections 3.3.1 and 3.3.2.

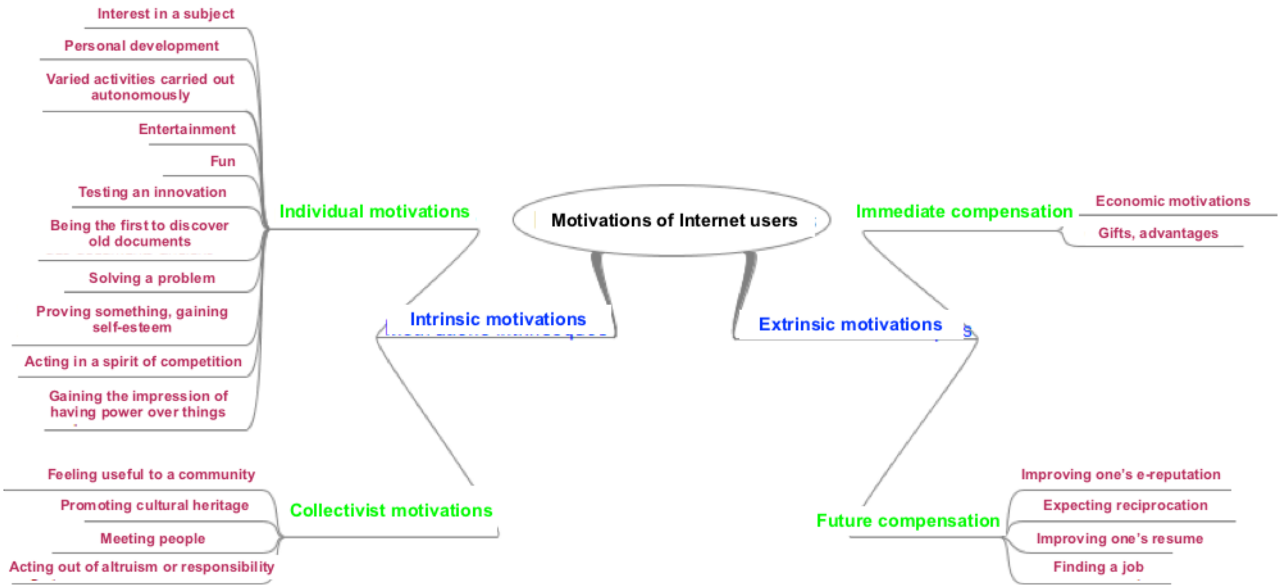

Motivations of Internet users
Intrinsic motivations
Extrinsic motivations
Individual motivations
Interest in a subject
Personal development
Varied activities carried out autonomously
Entertainment
Fun
Testing an innovation
Being the first to discover old documents
Solving a problem
Proving something, gaining self-esteem
Acting in a spirit of competition
Gaining the impression of having power over things
Collectivist motivations
Feeling useful to a community
Promoting cultural heritage
Meeting people
Acting out of altruism or responsibility
Immediate compensation
Economic motivations
Gifts, advantages
Future compensation
Improving one's e-reputation
Expecting reciprocation
Improving one's resume
Finding a job

Figure 3.6. *Taxonomy of the motivations of volunteers in a crowdsourcing project . For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

3.3.1. *Intrinsic motivations*

Among the intrinsic motivations, we have surveyed, in the literature, motivations linked to individual pleasure and collectivist motivations.

Individual motivations:

– interest in a particular science, the writings of a particular scientist or a discipline;

– for personal development, cultivation, learning, developing skills, satisfying the search for knowledge including self-knowledge (local history, genealogy);

– for amusement, to pass the time, prevent boredom and remain active;

– various activities performed in a free, flexible, autonomous and responsible way; with paid crowdsourcing, in particular, one can work freely when one wants, where one wants, as much as one wants, for whom one wants and choose the tasks that one wants to do;

– for pleasure, to have fun and play (gamification). Certain projects have even led to an actual addiction, generating working time of close to 60 h per week;

– to satisfy curiosity and test an innovative approach in the history of new technologies;

– to be the first to read historic manuscripts and have the opportunity to edit historical documents. Like an archaeologist who is the first to discover a relic that has been buried for a long time, being the first go do something important for a cultural document;

– to resolve an intellectual or technological problem. Carry out tasks that can still not be automated and performed by algorithms;

– in a spirit of competition, to prove what you can do as a challenge, including a collective one.

Consequently, numerous projects display the percentage of the project that remains to be completed in real time. They can also display the rankings of the best contributors of the week, month, year, of all time for a particular geographical area so that each one can hope to be in the rankings and encourage imitation. The projects also benefit from giving each person a dashboard with personalized statistics, giving medals and grades to contributors and, for crowdfunding projects, to rank the patrons in order of importance. On Amazon Mechanical Turk Marketplace, the rank of Master is earned by the workers who have contributed the most in quantity and quality according to the evaluations of what they deliver:

– to prove something, self-realization, improvement of self-esteem, to gain self-confidence, to feel effective, useful and competent. This motivation would be especially important for the unemployed;

– to have the impression that his or her opinion is being taken into account, that they are being consulted, that they have influence over things, that one can change things in the world, to leave one's mark [SHI 08], to be an author and actor, for vanity;

– not being a passive consumer of information, but an active producer of knowledge.

Social, community and collectivist motivations:

– feeling useful for a community, a group, for society, serving one's country by promoting its cultural heritage, working in the service of science, the general interest, the public good, acting for a cause, for values or for ideals (principalist motivations). Having the feeling of participating in a cause or in a movement which is bigger than we are. Digitalkoot, for example, explicitly appeal to patriotism: "Start saving… Finnish culture here". Many contributors to *OpenStreetMap* probably have the impression that they are fighting the hegemony of Google Maps. In the game puzzle fold.it, following the example of games where you have to save the world from invaders, we must try to penetrate the secret of proteins [GOO 11];

– participating in the free distribution, use and conservation of cultural heritage, participating in its reuse and in its promotion. In general, contributors do not want their work to be reused commercially by businesses. They also want to work for non-profit organizations such as libraries;

– possibility of meeting people, communicating, having interactions and being connected to a social network;

– doing something selfless, neither for personal benefit nor for money, in a spirit of altruism, sharing, generosity, charity and philanthropy;

– feeling indebted to services rendered by the site and feeling, in return, a duty to participate.

3.3.2. *Extrinsic motivations*

Immediate benefits:

– economic motivations (compensation for the work provided in the form of payments). This type of motivation could have a negative effect on other types of motivations;

– gifts, advantages. Some projects offer gifts (pens, books, t-shirts, etc.) or gift certificates from the library for digitization or printing on demand for those who have contributed the most. Others organize events and banquets in real life for the volunteers, or pay for them to visit the institution.

Future benefits:

– improving popularity on the web, improving one's e-reputation by showing up on the Internet as a volunteer in a cultural project, perform self-promotion, improve social status and satisfy a thirst for social recognition (especially for people who are unemployed). Benefit from the prestigious function of curator and to work for a famous institution. Certain projects individually thank their contributors with personalized e-mails and by public acknowledgements in the institution's written and oral communications on its website, its newsletter or on social networks. The scientific project Galaxy Zoo has thus added the names of Internet users to the list of authors of the scientific publications produced during the project. In folksonomy projects, the name of the person who added the keywords might also be cited. Finally, for crowdfunding projects and digitization on demand, the name of the patrons and a link to their websites must be indicated in order to allow them a return on their investment in terms of web traffic if a book generates a large number of visits (following the Google Adwords model). On YouTube, contributors are increasingly active as their videos generate web traffic [HUB 09];

– searching for reciprocity (one will receive more help on the Internet more easily if one has already helped others him or herself, "I do it because I would like it if someone did it for me".);

– finding a job or undergoing a career change because of this self-advertising;

– personal development for a career change.

3.3.3. *The opposition between intrinsic and extrinsic motivations*

Among all of these motivations, we can distinguish individualistic motivations (increasing one's own wellbeing), altruistic motivations (to increase the well-being of fellow human beings), collectivist motivations (increasing the well-being of the group) and motivations based on principle (defending a moral principle such as freedom, equality, brotherhood or justice).

A study of the motivations of volunteers working on cultural crowdsourcing projects [BRA 10], created from data collected through instant messaging from seventeen people in March, April and October 2008, showed that the intrinsic motivations (pleasure, amusement, problem solving, improving skills, addiction) predominated over more extrinsic motivations (money, professional opportunities, love of the community). The altruism of the contributors would, however, be debatable for certain projects. Thus, for the ACM Digital Library project, the majority of corrections carried out online on bibliographic references were simply the work of the authors themselves [BAI 12].

According to another survey conducted by Dunn and Hedges [DUN 12], 79% of active contributors act out of motivations simultaneously for themselves and for others. Out of 59 people, 24 claimed to be acting out of interest in the subject, three to help others learn, two to contribute to science, two to experience crowdsourcing, one to be in involved in volunteering and one for the novelty. Only one estimated that a computer algorithm could have been used in place of human labor. According to an analysis of 207 messages on the Galaxy Zoo project forum [RAD 10], he found that the main motivation was astronomy (39%), followed by the desire to contribute (13%) then an interest in the immensity of the universe (11%).

Acar and Van Den Enden [ACA 11] studied the impact of bonuses on workers mostly driven by intrinsic motivations and seemed to find a negative effect of gratifications on this type of person. Furthermore, the quality of the data produced would be improved with calling on intrinsic motivations as Rogstadius *et al.* [ROG 11], who compared the quality of the data produced for free out of intrinsic motivations with that of data produced against compensation from extrinsic motivations, suggest.

Despite the interest in intrinsic motivations from cultural projects, the experience of the TROVE project reminds us, however, not to ignore extrinsic motivations. Indeed, during the first 6 months of the TROVE project, half of contributions were anonymous and the fruit of more intrinsic motivations (personal interest, altruism). Six months after the launch of the project, only 20% of contributions were still anonymous. Volunteers therefore probably also have a need for recognition. This is the reason why extrinsic motivations were then developed in the form of statistical rankings by the heads of the TROVE project.

3.3.4. *The specific motivation of gamification projects*

The motivations mobilized specifically by gamification projects seem to be the following:

– personal development (acquisition of skills, problem solving);

– rewards (money, prizes, promotions, recognition, responsibilities);

– amusement and distraction;

– information (on the progress of the project, the size of their own contribution).

According to [MCC 12], male players had more of a tendency to evaluate their performances than female players who are more attracted by the interpersonal character of the games. In general, women were more attracted to management, puzzle, combat and adventure games, and for their part, the male gender has a preference for sports, shooting, strategy or role-playing games.

According to [DUN 12], gamification can sometimes also be an obstacle for certain users who want to commit or who are interested in a subject, since the development of knowledge can be lower. Furthermore, certain players risk producing a large amount of low-quality work solely so they can be highly ranked. For example, the Old Weather Project, which transcribes manuscript pages of 19th Century ship’s logs containing meteorological observations, some risk neglecting quality to go more quickly from the rank of cadet to that of lieutenant then to that of ship’s captain or also to be able to keep their title of captain. Others risk losing motivation and give up trying to measure themselves against players who are too highly ranked and too difficult to dethrone [EVE 13]. Consequently, without inciting an addiction to the microtasks offered, but in order to increase the motivation of players, Von Ahn and Dabbish [VON 08] suggest showing, along with the very best players, the very best players for the month, for the week, of the day, etc., in order to encourage even more the participation of players who over a week can hope to end up on the podium. In the same vein, Ridge [RID 11] indicates that many players are driven both by immediate and local desire to defeat the player located just above in the rankings, but also by the long-term goal to beat the highest score. Under these conditions, he suggests that the list of the highest scores cannot only be displayed by the hour, day, week, month, year, or for all time, but also by town, region, country and continent. By crossing these two variables, one could display, for example, the list of the best players by country and per month or that of the best players by town and by year. To the extent that the players seem to react differently to different types of goals, it would be thus possible to display the type of goal, which corresponds best to each person's personality.

The possibility of the best players winning a gift or a sum of money could also be an excellent way to increase their contributions. One could easily imagine that such a gain would be much more attractive than games of chance since it would be actually possible to win through tenacity. The value produced by all of the players serves to generate gain. Another model could be to pay the players for the level of their contribution.

3.3.5. *Crowdsourcing and rewards*

In the majority of crowdsourcing projects and not exclusively in gamification projects, contributors are ranked according to their contribution, just as they are in video games. Thus, Internet users, in their own spaces, generally have access to their statistics and the lists of documents on which they worked. This can be very beneficial for their self-promotion, their e-reputation and during a job search. Thus, the contributors to the Galaxy Zoo project were thanked and acknowledged in the articles resulting from the project, and a super contributor was invited to attend a prestigious open conference organized around the Transcribe Bentham project. However, in addition to social rewards, symbolic, or material and in-kind rewards are offered by certain projects such as Archive Cooperative Engine for Correction of Extracted Text (CONCERT) and TROVE, and actual financial payments are granted to Internet users working in the service of libraries on the Amazon Mechanical Turk Marketplace or on CrowdFlower. In this situation, we are talking about paid crowdsourcing.

It is very important to attempt to gain the loyalty of participants in order to obtain data on skilled participants, and consequently higher quality data. To do this, we can display the list of the largest contributors, display the name of the contributor for each contribution, promote a particular contributor by highlighting his or her biography in a newsletter, acknowledging them individually, reward them with certificates, training, recognized diploma courses, gifts, subscriptions, books, organize outings, events or participation in the analysis of the results [BAU 10].

Regarding, for example, the Australian ArcHIVE, the compensation takes the form of a symbolic payment allowing the exchange of points won against facsimiles (Print on Demand), objects, bookmarks or posters. This model has also been adopted by the cultural crowdsourcing site velehanden.nl (consulted June 23, 2016), which allows the conversion of points accumulated into gifts, services or financial compensation [DJU 13]. Regarding the TROVE project, Rose Holley has mentioned the possibility of offering gifts, t-shirts, books, certificates, training, public acknowledgment ceremonies on the web, on social networks, in bulletins or special visits to the National Library of Australia collections for the best contributors [HOL 09a]. The literature also mentions the invitation to meet the head of cartography at the institution as part of the British Library Georeferencer gamification project [DUN 12], but also MP3 players, free products, access to advanced functions on a platform [BIE 15] or even small financial rewards offered to volunteers [ROU 10], or finally, Amazon gift certificates [BIR 12]. Finally, regarding crowdfunding projects such as Numalire, the return on investment for a patron or an Internet user can be measured in terms of web traffic generated by the books which it has allowed to be digitized, and the publicity gained for his or her name or the name of his or her business or institution.

According to a study reported by Ipeirotis and Paritosh [IPE 11], money does not have an impact on the quality of the data produced but generally has impact on participation. Nevertheless, the act of going from an intrinsic motivation to a more extrinsic incentive can also provoke negative effects.

Rogstadius *et al.* [ROG 11] estimate, for example, that a low payment amount could have a less positive effect than an absence of payment. Thus, exterior pressure, and in particular extrinsic rewards could have a negative effect on the intrinsic motivations that underlie, for example, gamification. Paying Internet users thus has the effect of making them feel that they are losing their autonomy and their freedom and paradoxically decreases their desire to play [HAM 14].

Thus, as Groh [GRO 12] relates, experiments have shown that if they are paid for drawing, children might draw more, but these more numerous drawings will be of lower quality, and that the children will have lost their desire to draw, especially if they subsequently stop being paid. In the same way, as Rogstadius *et al.* [ROG 11] report, a 1975 experiment by Edward L. Deci showed that students who were paid for playing puzzle games had also lost all interest in this activity once they stopped being paid. In his book, *Homo Economicus: The (Lost) Prophet of Modern Times*, published in 2012, Daniel Cohen relates how the director of a blood transfusion center decided to offer a gift to blood donors and that this action had the paradoxical effect on significantly decreasing the amount. As the author says:

> "If it's no longer a matter of helping others but earning money, their participation changes in nature. A different sphere of their brain is being called upon. The moral person leaves the room when *Homo economicus* enters".

Michel Bauwens uses the term *crowding out* for this phenomenon [BAU 15]. The human being is a complex being endowed with free will and does not act exclusively out of love for the carrot or fear of the stick.

3.3.6. *Other theories on motivation*

More generally, according to Maslow's theory, needs are distributed hierarchically in a pyramid in the following way with vital needs at the bottom (physiological needs for existence, safety, belonging) and higher needs at the top (need for relationship, esteem, power, progress, realization).

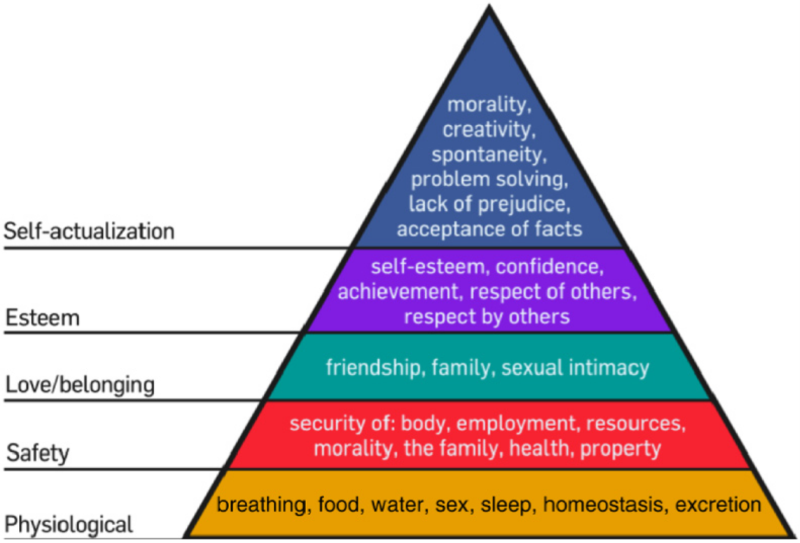

Self-actualization
morality, creativity, spontaneity, problem solving, lack of prejudice, acceptance of facts
Esteem
self-esteem, confidence, achievement, respect of others, respect by others
Love/belonging
friendship, family, sexual intimacy
Safety
security of: body, employment, resources, morality, the family, health, property
Physiological
breathing, food, water, sex, sleep, homeostasis, excretion

Figure 3.7. *Maslow's Hierarchy of needs, By user: Factoryjoe (Mazlow's Hierarchy of Needs.svg) [CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons (consulted October 4, 2017). For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

When the primary needs have been satisfied, new needs, higher ones, appear. However, on the other hand, these cannot exist unless basic needs are not already satisfied. Under these conditions, it is important not to overestimate the reward to the detriment of interest in the work itself. Not taking into account intrinsic needs could even dehumanize the work, a salary not really being a motivating factor, but rather a factor in satisfaction. It is the same for projects that call upon crowds of Internet users.

Gouil [GOU 14] cites the work of Stéphane Debove who estimates that biological reasons such as the parental instinct, the need to favor transmission of our relatives' genes, the need to increase our group's chances of survival through collaboration, the possibility of improving our reputation, finding a partner and thus transmiting our genetic material could explain why Internet users cooperate so much on the web. In a less academic, but very effective, way Simon Chignard relates on his blog[2] that when automobile drivers flash their beams at other drivers in order to warn them of the presence of police radar, they are acting both in expectation of reciprocity, or because they feel

[2] See: https://donneesouvertes.info.

indebted to others for having already been warned, due to opposition to police officers, or by a feeling of solidarity with the community of drivers.

Many other classifications of motivations exist. Therefore, Herzberg identifies five factors for satisfaction: achievement, the recognition of the achievement, the work itself, responsibility and professional advancement. For their part, the theorists Porter and Lawle feel that action must be motivated by the following three factors: interest and the challenge of the action, the consideration of the actor's social relationships, the capacity of the actor to lead the action. For McClelland, motivation is influenced by the variety of activities, the tasks that one can carry out completely and for which one can claim authorship, the meaning of the tasks, autonomy, the possibility of deciding and, finally, the return that these tasks can provide. Individuals are also driven by diverse types of motivations. Some need self-realization, others power, and finally others need membership or affiliation. In the opinion of Douglas McGregor, there are two types of conceptions: those who think that humans have a natural aversion to work and flee any type of responsibility. It would be, consequently, necessary to control and educate workers to obtain work from them and to use the method of the carrot and the stick to move them forward from fear of punishment and desire for a reward. This theory, known as "theory X" is shared both by the great capitalist Ford and also possibly by certain Marxists such as Paul Lafargue with his "right to be lazy". Others have a diametrically opposed theory, "theory Y". According to them, individuals naturally love to work, deriving satisfaction and pleasure and seeking out responsibilities. Under these conditions, work is liberating and fulfilling, it allows social realization and development just like with hobbies and leisure. Consequently, it is necessary to favor trust, responsibility, autonomy, freedom, sense of initiative and creativity of employees in order to motivate them and obtain an optimal result from them. It is this type of conception of motivations that crowdsourcing is founded upon. The motivation of volunteers will depend on the variety of their tasks, their autonomy, their responsibilities, information and feedback on these tasks.

3.3.7. *The motivations of cultural institutions and the prerequisites for launching a crowdsourcing project*

According to a 2010 survey reported by Thuan *et al.* [THU 13], 10% of businesses had deployed a crowdsourcing strategy. However, as Alam and

Campbell [ALA 13a] emphasize, few studies are interested in the motivations of organizations and institutions in crowdsourcing. The article emphasizes the fact that the motivations that drive institutions to use crowdsourcing are the same as those that lead them to outsource. It involves, in particular, reducing costs and improving the result/cost ratio [LEB 15]. Regarding libraries in particular, it will involve decreasing costs in the context of tightened budget, speeding up projects for which they do not have sufficient human or financial resources, or launching projects that could not have previously been carried out for these reasons. It also involves gaining access to skills and knowledge not available internally and going beyond those of a limited team, to benefit from the skills of scholars and researchers, to better adapt its services to the needs and to better educate the general public about the activities of professionals. To resolve problems impossible to resolve without crowdsourcing: to improve the quality of the data and index collections or enrich them including new types of information to remain technologically relevant in a rapidly changing society; to be innovative; maintain leadership; to increase recognition; to use budgets, previously used to pay for work in countries where labor is cheap, in a more useful and ethical way and, last but not least, to seek out new types of relationships with users.

Indeed, beyond the need to call upon Internet users to capture the free labor force on the web, to subcontract tasks that it no longer has the means to finance, or even to initiate projects that they could have never hoped to develop without the help of Internet users, crowdsourcing is also, for some people, above all, the means of extending the mission of cultural institutions, to engage the public more in the service of themes and collections, by involving it in the conservation of cultural heritage and the public memory in order to produce new knowledge. It also makes it possible to change the public's opinion of the museums and libraries that are currently not always considered playful and fun [BIR 12]. The creation of a new community, outside of the walls, attached to the institution and/or its collections thus also becomes a purpose in itself. It will involve, for the library, building and running a real community of Internet users around its digitized collections. The use of digitized cultural heritage by Internet users will therefore be less superficial, less passive and could lead to real research. Instead of consuming information, they can become producers of information themselves. Instead of asking people to work for the library, it will instead involve the possibility of participating in the enrichment of common heritage. On his blog, Trevor

Owen[3] believes that crowdsourcing, in its best form, does not consist of having users work, but rather offering them the possibility of participating in collective public memory.

It seems in any case that there are two clearly different conceptions of the interest in using crowdsourcing for institutions, two conceptions which are moreover not necessarily in opposition. Some seem to focus on the interest of institutions in terms of costs and others on the public's commitment to the collections. The question of costs must not be dismissed as evidence of a profit-driven vision that lacks merit. On the contrary, institutions must more modestly recognize that the investment of Internet users is necessary, or even vital, to them. They cannot believe that they are just entertaining Internet users by allowing them to express themselves. They cannot remain content with conducting crowdsourcing following a simple logic of institutional communication around a fashionable subject and never reintegrating and reusing the data produced by Internet users, as is still unfortunately too often the case. "It would indeed be a shame to use the potential of the social web only 'cosmetically', without it truly benefitting the description of collections and the library's search interface" [MOI 13b].

Whatever may be said, the principle force for crowdsourcing remains the decrease in costs, and obtaining work abilities or skills that are not available internally. As with any outsourcing, the payment (when it does not involve volunteer work) is based on results and not on the time spent working, which presents a certain advantage compared to salaried employees.

Beyond the motivations of institutions, some conditions are necessary for the deployment of crowdsourcing by businesses. Thuan *et al.* [THU 13] and Crowston and Prestopnik [CRO 13] have thus identified the type of tasks (can be completed through the Internet, non-confidential, able to be carried out independently and requiring little interaction and communication, able to be divided into microtasks and able to be completed by non-experts), the type of workforce (a larger and more diverse crowd via crowdsourcing when the human resources and skills available internally are not enough, presence of already-existing communities of enthusiasts), the type of management (the budget they have access to is insufficient and requires resorting to crowdsourcing, the presence of human resources who have experienced or

[3] See: http://www.trevorowens.org (consulted June 23, 2016).

are experts in crowdsourcing, the level of quality required) and finally, the work environment (internal or external platform).

3.4. Sociology of the contributors and community management

In the conceptual introduction (Chapter 1), we have already mentioned the sociology of the contributors to crowdsourcing projects in a general way. Here, we will deal more specifically with users in the area of digital libraries in particular and in light of the projects that we have analyzed.

3.4.1. *Sociology of contributors*

Here, in Table 3.6, is a summary of the sociological information collected as part of the overview of projects.

	Gender	Age	Social status
Crowdfunding (Numalire)	Men (70%)		High
Crowdfunding (eBooks on Demand)	Men in the majority		High
OCR Correction (TROVE)	Women (70%)	Recent graduates searching for employment and retirees, older than 50 years old (65%)	Graduates
Manuscript transcription (Transcribe Bentham)	Women (more than two-thirds)	Retirees Recent graduates	High
Tagging (VeleHanden)	Men more numerous	More than 50 years category is the most numerous	
Gamification (Digitalkoot)	Women are as numerous, but more dedicated (54% of tasks), but the four largest contributors are men	Between 25 and 44 years old	
Gamification (Art Collector)	Men 10% more numerous	25–34 years old	
Gamification (Museum Games)	Women more numerous	Thirty-somethings	

Paid crowdsourcing (Amazon Mechanical Turk Marketplace)	American women Indian men	Relatively young	High
Crowdsourcing survey [DUN 12]	Among the respondents, 58% men and 42% women	The majority in the 35–45 age range	
Crowdsourcing survey [MCK 13]		Domination of preretirees (56–65 age range)	

Table 3.6. *Data collected in the literature about the sociology of the contributors to different projects*

If we encounter the rather significant gender differences for each project, it remains difficult to draw correlations according to the types of projects outside of digitization on demand through crowdfunding, which seems to attract more men. In general, crowdfunding attracts more high-income men [DAU 14]. Gamification could also attract more men while explicit crowdsourcing attracts more women.

The report from the Wikimedia association [BEU 14] shows that nine out of ten editors on Wikipedia are men and that this proportion is 97% for the Indian version of Wikipedia. Women prefer to spend more of their time on social networks such as Facebook (in the United States, their level of participation in Facebook represents 71% of the total); they also tend to have less free time than men and do not like the sometimes somewhat virulent, polemic and aggressive tone of the discussions on Wikipedia.

Regarding education level or social status, when this information was able to be collected, it is people who have a high educational level who seem to be in the majority among the contributors.

Regarding the age of the contributors, we can observe a clear dominance of young people in gamification and paid crowdsourcing projects while OCR correction or the transcription of manuscripts seem to attract older volunteers. This observation is in line with [DAU 14] who reports that 44% of 12- to 17-year-olds contribute on the Web.

According to [HOL 09c], the majority of volunteers in the TROVE project are retirees, but there are also young graduates searching for work or the unemployed or salaried employees on medical leave or vacation who contribute. And, regarding positions of responsibility such as moderation, they are instead taken charge of by full-time 30- or 40-something salaried workers.

The dominance of retirees passionate about genealogy and local history in traditional cultural crowdsourcing poses the question of its longevity, since there is nothing to indicate that future generations of retirees will have the same areas of interest [AYR 13], nor do they have as much free time.

3.4.2. *Crowdsourcing or community sourcing?*

The data collected as part of the overview of projects revealed that, for the majority of projects, the majority of data produced are the work of a small, well-defined minority of participants and not an anonymous crowd, and this is despite the fact that these projects are addressed in theory to an unlimited number of Internet users.

These observations are in keeping with that which is reported in the literature and especially by Brabham [BRA 12]. Thus, 80% of the work is carried out by barely 10% of the most active volunteers. Some are likely to dedicate so much time to it that they do it as full-time work or even experience a sort of addiction to this activity.

Therefore, we should finally rather talk about community sourcing (or community-sourcing) or nichesourcing (or niche-sourcing) rather than crowdsourcing involving these participative digital library projects. Nichesourcing could even be the future of crowdsourcing [DEB 12]. By recruiting small communities of expert amateurs with a wide diversity of profiles, journeys and points of view, we also obtain groups capable of making better, more intelligent and wiser decisions [SUR 04]. With community sourcing, instead of entrusting repetitive microtasks or atomic tasks to a faceless crowd, to develop communities of practice and interest, we assemble peers that have an identity, affinities and above all common goals. The regular exchanges between its members progressively foster social trust and increase each person’s reputation.

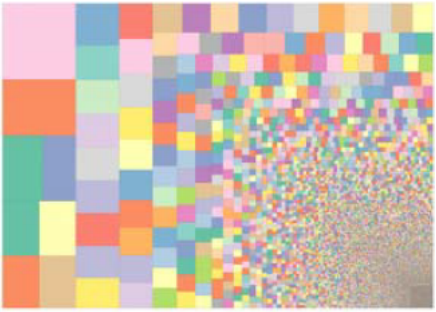


Figure 3.8. *Diagram showing that a handful of Internet users are the source of the majority of contributions, from Brumfield, manuscripttranscription.blogspot.fr, 2013*[4]*. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

Consequently, the Rijksmuseum found that it instead needed amateurs, experts, the self-taught and retired professionals rather than a true crowd of Internet users and focused on community sourcing [DEB 12]. In the same way, the MarineLives (ML) project thus called upon not to just anyone, but participants who agreed to work at least 3 h per week and who committed to 14 weeks minimum [DUN 12].

3.4.3. *The work of professionals on these projects and community management*

As Ellis [ELL 14] emphasizes, regarding crowdsourcing as just a source of free labor is a serious error. This is somewhat similar to forgetting that even if you acquire a pet for free, it still must be fed, trained, walked, cared for, etc. Without true management of the crowd, the obtained result risks being dramatic. The free labor of Internet users will be largely made up for by other costs. Oomen *et al.*'s [OOM 10] study, which reports on the experiment of the Netherlands Institute for Sound and Vision with the game Waisda, also underlines the importance of supporting and supervising volunteers and laments that this factor has sometimes been underestimated by project managers. A report published by OCLC [SMI 12], therefore, recommends the recruiting of a community manager. A study conducted in this same framework shows

[4] In this diagram, already presented in the overview of projects, each square represents a contributor. The size of each square is proportional to the quantity of his or her contributions. We thus observe that a minority of volunteers are the source of the majority of contributions. This observation is verified by all of the projects using volunteer work.

what the activities of the teams which support crowdsourcing projects are (see Figure 3.9).

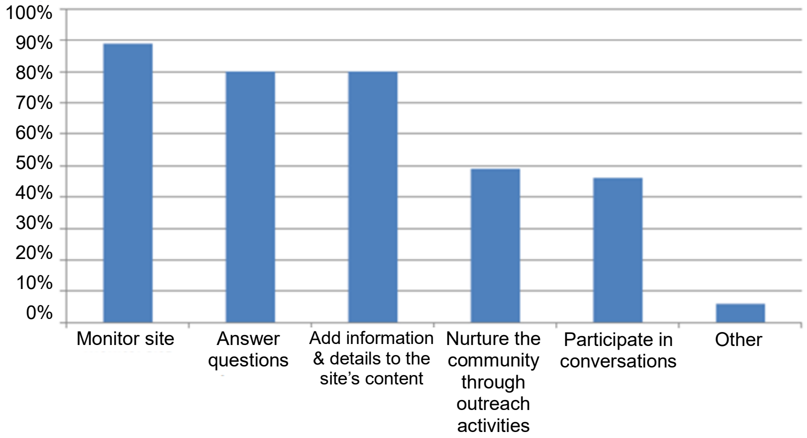


Figure 3.9. *Distribution of staff activities in management of crowdsourcing projects, from [SMI 11]*

In addition to the platform's administration and configuration, according to the response to questions from volunteers, participation in discussions, the addition of information and news, training of the volunteers, we could add, in order to describe a more complete job description of community manager, moderation of the contributions and support forums, quality control, statistical data collection, the development of functions, reintegration of the data produced, project management, communication, recruitment, motivation and conservation of volunteers, the writing of blogs, manuals, guides, tutorials, frequently asked questions (FAQ), contextual help, the definition of rules, creating screencasts, demonstration videos, the development of "sandboxes", managing a hotline, a helpdesk [ZAS 14] and other activities, which have a cost which risks mitigating, canceling out, or even surpassing the free labor collected through the project.

According to the survey already mentioned, it is often professionals who dedicate a part of their time to community management or full-time professionals on crowdsourcing projects, less often professionals not specializing in the area of Internet technologies and even less often, volunteers trained by professionals who also participate in writing procedures, rules and manuals. For 23% of respondents to this survey, the volunteers trained by professionals play a role in the management of crowdsourcing websites.

In order to help Internet users be trained to manage themselves, tools can be offered to contributors and also written with them collaboratively. Sometimes, volunteers can even be tasked with moderating contributions and coordinating the work of other volunteers, because of participative tools such as *wikis* [HOL 09c]. Volunteers must be able to easily train themselves with the help of tutorials, screencast videos, forums on which they must be able to ask for help. All of the contributors have all of the characteristics of remote workers. They must, consequently, be able to be trained remotely and the traditional telecommuting and e-learning devices must therefore be mobilized for them.

The time that professionals spend on platforms seems to be very variable from one institution to the other as shown in Figure 3.10.

The time spent on crowdsourcing projects is distributed among the respondents to the survey in the following way.

With the development of crowdsourcing in libraries, the profession of librarian could undergo an evolution and go from cataloguer-indexer to community manager.

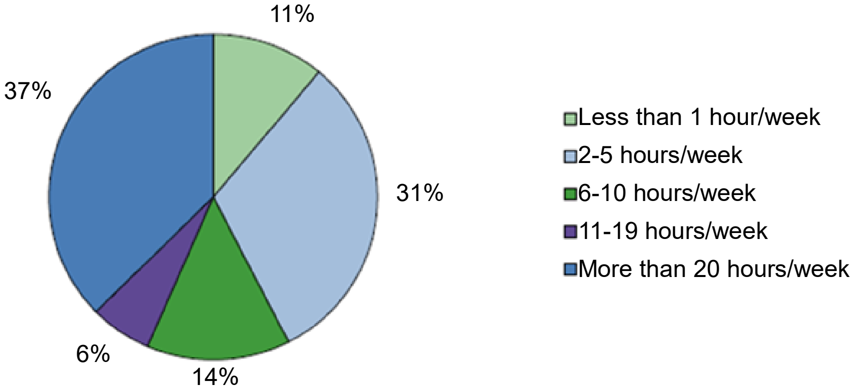
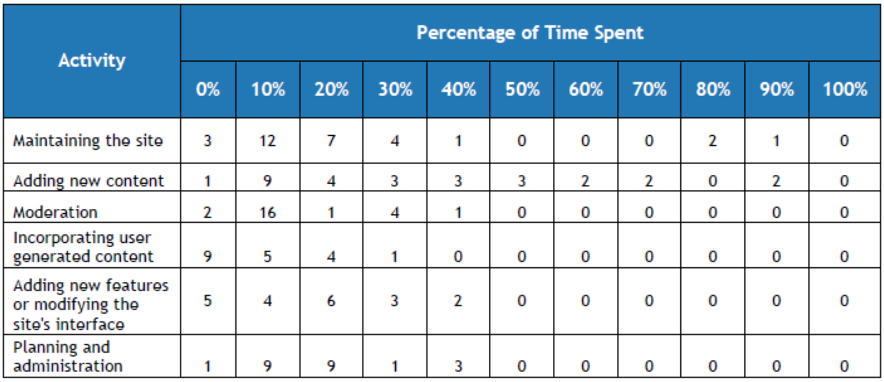


Figure 3.10. *The working time of crowdsourcing project staff, from [SMI 11]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*



Activity	Percentage of Time Spent										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Maintaining the site	3	12	7	4	1	0	0	0	2	1	0
Adding new content	1	9	4	3	3	3	2	2	0	2	0
Moderation	2	16	1	4	1	0	0	0	0	0	0
Incorporating user generated content	9	5	4	1	0	0	0	0	0	0	0
Adding new features or modifying the site's interface	5	4	6	3	2	0	0	0	0	0	0
Planning and administration	1	9	9	1	3	0	0	0	0	0	0

Table 3.7. *Distribution of the working time of crowdsourcing staff according to activities and missions, from [SMI 11]*

As we saw in the chapter about motivations, there can be a wide variety of different motivations different from one individual to the other. The community manager must know how to play on its various motives. In order to unite the contributors in their diversity, and in order to create loyalty, it is necessary for Internet users to be able to easily acquire the digitized cultural heritage, by having access according to traditional metadata criteria (title, author, date, subject, geographical area, etc.), they must be able to easily choose the type of documents to which they want to contribute (eras, themes, authors, etc.), but also have access to documents to be processed depending on their levels of difficulty, by degree of progress, or simply randomly by document.

The participation of Internet users must also be regularly maintained by periodically adding new content to be processed; the site must be editorialized in order to increase the activity of volunteers. This periodic uploading of new content is preferable to posting everything at the same time, which could have the effect of discouraging volunteers. Thus, 59% of cultural crowdsourcing sites studied as part of Smith-Yoshimura *et al.*'s [SMI 11b] survey claimed that they uploaded new content at least once a week.

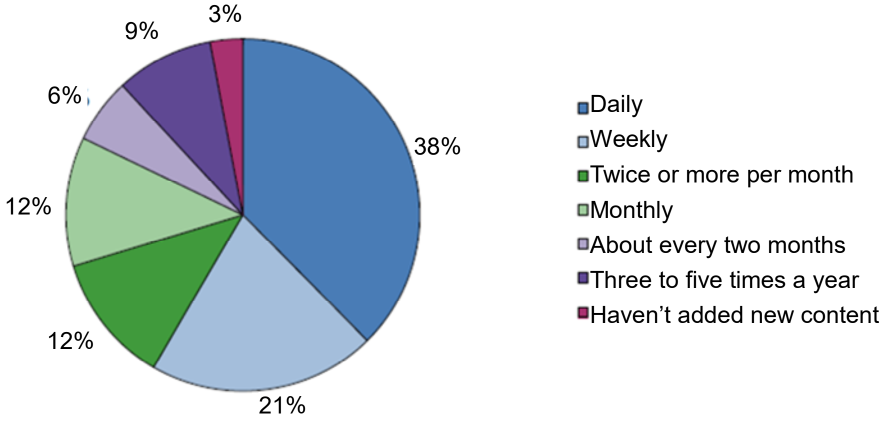


Figure 3.11. *Frequency with which sites put new content online, from [SMI 11b]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

In order to better “keep the flame going”, it is also necessary to take advantage of current events, special occasions and historical anniversaries to encourage volunteers to contribute more.

Finally, according to [MOI 13a], some projects offer a structured and hierarchical organization of contributors into communities. This is the case of the CONCERT, which offers activities depending on the experience and skills of Internet users, of Monasterium that involve expert administrators or Transcribe Bentham that ranks contributors according to the quantity of their contributions. Other projects do not offer any structure to the community of contributors such as Ancient Lives, ArcHIVE, Digitalkoot, Do it yourself history, TROVE or What’s on the menu?

3.5. The question of the quality of the contributions

As has been discussed at length in the conceptual chapter, a central objection of the opponents to setting up crowdsourcing involves the quality of the contributions.

Sturgeon’s law therefore claims that 90% of ideas collected from crowds are mediocre and that only 10% of them can be of a quality equal to that of specialists [ROT 16]. For some, the free work must necessarily be substandard work. Moreover, in the USA, in the December 27, 2010 *New York Times*, Patricia Cohen reported some less than reassuring statements on this subject from Daniel Stowell, head of the Papers of Abraham Lincoln. According to Daniel Stowell, the number of errors present in the contributions would require finally spending more time and money on their correction than if they used work traditionally produced by professionals.

We will see, however, that there are systems that make it possible to ensure a high quality of the contributions and to evaluate their quality and that there are studies that have sought to compare the quality of the data produced by amateurs and professionals. Finally, we will address the subject of reintegration of the data produced by Internet users and the legal status of these data.

3.5.1. *Systems for evaluating and moderation of contributions*

The report commissioned in 2012 by OCLC [SMI 12] on the subject of social metadata for libraries, archives and museums, recommends not worrying too much about spam or vandalism. Spam can, for example, to be filtered automatically with the help of CAPTCHA that can verify the human origin of the contributions. It also remains possible to require Internet users to log in to contribute in order to be able to limit and spot any wrongdoing or register their IP addresses and ask Internet users to monitor the quality of the contributions themselves, as Wikipedia did to protect itself from vandalism in conjunction with the use of robots. Going even further than the OCLC report, Holley [HOL 09a] recommends not assuming that everything will go badly and wasting valuable time setting up systems to prevent vandalism. She recommends trusting Internet users, assuming that they will do their best, and providing them with a maximum amount of freedom. She observes that the Australian project TROVE is founded on these principles of trust and have never from suffered from vandalism. Others feel that the appearance of vandalism in a project means that it has left the prototype phase to enter into a maturity phase.

However, beyond the simple question of vandalism, there is the question the quality of the contributions; it is moreover regularly cited by those expressing mistrust of systems with participative mechanisms. As [MOI 13a] emphasizes, in order to guarantee high enough quality of contributions, volunteers must be trained, assisted and evaluated, and they must be offered tasks according to their skills, compare their contributions and control its quality.

Starting from the overview of projects, not only in the literature ([QUI 11] and [KLE 14]), but also blogs such as that of the developer Ben W. Brumfield (manuscripttranscription.blogspot.fr, consulted on June 26, 2016), we propose the following typology for quality control systems for the contributions.

There are also quality control systems:

– no quality control and trust in the self-regulation of the participants;

– quality control by experts, professionals, librarians – an effective method, but a very thankless and very expensive:

- revision over a determined or indeterminate period of time by experts who lock in and publish the contributions of transcription has been achieved (examples: Transcribe Bentham, Scripto, Do it yourself history, Monasterium, What’s on the menu);

- deliberate insertion of mistakes and traps into documents in order to verify that the quality control has been done and in order to ensure vigilance and the quality of the contributor. This method was notably used by an overview project that relied on paid crowdsourcing for the transcription of manuscripts;

- use of benevolence tests, used by the CONCERT;

– quality control by other volunteers – by the community of volunteers, the way Wikipedia does, or by submitting contributions to a vote:

- quality control through division of labor. In the string of microtasks, the previous task is thus controlled and verified by the following operator. Wikisource’s workflow uses this method, for example the same person cannot simultaneously be the one who validates a line and the one who validates a page. Someone who corrects a page will see his or her work evaluated by the person who validates a page, who will see his or her work evaluated by another person who validates a text;

- quality control of volunteers by other volunteers. A group of Internet users verifies the work of a previous group. In digitization markets, for example, it happens that a service provider performs quality control on an original service provider;

- quality control through comparison of entries. This method is used in particular by companies working in human OCR correction, in Madagascar, for example. The same text is transcribed by two operators then the differences in entries are compared in order to obtain a transcribed text of optimal quality. The process used by certain crowdsourcing projects is somewhat similar. It is one of the most effective and most tested methods for guaranteeing the quality of the contributions. Within this method, several submethods exist:

- presenting the same word to be corrected to two different contributors. In case of divergence, it is either an expert or a major contributor who decides, or the word is presented again to two new contributors. It is, for example, this method that is used by the gamification site Digitalkoot;

- present the same word to be corrected to several different contributors (three minimum). The majority prevails. It is, for example, this method that is used by reCAPTCHA for the Google Books project;

- quality control by engines, algorithms, statistical filtering methods. This method is notably used by Wikipedia.

In concrete terms, according to the survey conducted by OCLC [SMI 11], 75% of respondents (27 out of 36) say they have moderated contributions; 36% of respondents approve each contribution before posting it and 50% of respondents can edit the contributions. This result somewhat surprised the experts in charge of conducting this study. In fact, this systematic control activity can be very time consuming and expensive. Moreover, 53% of respondents claim that their website requires an identification, 36% use a CAPTCHA-type system, 36% use the contributor's e-mail address. In only 31% of cases, no identification is required. That being said, spam was a problem for only 6% of respondents. 69% did not encounter this problem and 25% encountered it only occasionally. Only 36% of respondents (13 responses) had already encountered malicious users attempting to add inappropriate contributions.

Among the strategies used by institutions to guarantee the quality of the contributions, the OCLC survey identified the following:

– the institution retains the right to modify, reuse or delete content generated by the user without prior notice (57%);

– users who violate the policy can be blocked from the site (31%);

– the ownership of the content generated by the user is retained by the site/institution (31%);

– a chart showing the project’s guidelines and its mode of operation (14%);

– no specific policy (11%);

– trusted users can contribute without moderation (whitelist) (9%).

According to [MOI 13a], certain sites require mandatory prior authentication before being able to contribute such as CONCERT, Monasterium, Transcribe Bentham and Digitalkoot (authentication via Facebook) while for others authentication is optional such as Do it yourself history, TROVE, What’s on the menu? or Wikisource. Some institutions provide total freedom to Internet users who want to contribute and they do not even have to sign up, while others supervise or control them and even have them pass paleography tests beforehand.

The documents are sometimes also classified according to their difficulties and assigned according to the skills of the contributors who have, themselves, been evaluated. Assessment tests are sometimes even given to volunteers in order to determine their level and assign them tasks adapted to their faculties.

Beyond classic validation by experts, recording the history of each modification and the possible restoration of a previous version is a useful means for ensuring the quality of the contributions and avoid vandalism. This is how, for example, TROVE or Wikisource functions. Internet users must also be able to easily point out errors. Google Books thus allows, for example, its users to indicate that a particular document has been badly digitized.

Another effective way to guarantee the quality of the work within paid crowdsourcing, in particular, can be to design tasks in such a way that it is not easier to cheat than to actually do the task or to play on the reputation of the contributors whose work is evaluated the same way eBay sellers are [QUI 11]. Consequently, on the Amazon Mechanical Turk Marketplace, workers’ statistics are visible and it is possible to know the number of tasks validated and rejected for each one. It is also possible to punish malicious Internet users by banning them (account blocked, contributions deleted), which can harm their e-reputation and even have an impact on their social and professional lives [DUN 12]. This is known as public shaming.

Finally, Eickhoff *et al.* [EIC 12] observes that paid crowdsourcing platforms are rife with malicious workers who attempt to maximize their profits dishonestly. He observes that they are especially common among certain nationalities and therefore suggests restricting the offer of work to some other nationalities among which there are fewer dishonest workers. However, this could simultaneously pose ethical and possibly legal problems.

3.5.2. *Comparison between the quality of the data produced by amateurs and that produced by professionals*

Theorists of the wisdom of crowds, such as James Surowiecki, think, as we have seen in the conceptual chapter, that the diversity of the profiles contained in a crowd have a tendency to provide much better results than the opinion of the top specialists in a field when it comes to decisions. The "law of large numbers" makes it possible in any case, in the field of citizen science, to neutralize individual errors within the mass of accurate data provided by crowds [BOE 12]. We are also aware of studies highlighted by Wikipedia[5] which show a quality equal to or greater than the participative encyclopedia compared to traditional encyclopedias with a limited review panel and which finally have more chances to let errors go by than in an encyclopedia where the entire world can correct them. In the area of participative sciences, players of the game fold.it could also in many cases produce better results involving the proteins that the Rosetta program, according to certain authors [GOO 11]. Despite these arguments and despite all of the mechanisms that we have mentioned in the previous chapter and which make it possible to guarantee the quality of the data produced, comparison of the result obtained by amateurs with that of professionals can be legitimate. It has moreover been the subject of multiple studies.

Thus, a university study [THO 12] sought to compare the quality of the indexing of images obtained by amateurs via crowdsourcing mechanisms in the form of gamification and by professionals via a more traditional method. It appears that crowdsourcing has the tendency to privilege, subjectively, the content, that is to say what is represented in the image, while professionals are more objectively attached to the form and the objects. Another study

[5] Giles J., "Internet Encyclopaedias Go Head to Head", *Nature*, vol. 438, pp. 900–901, 2005.

J. Trant in 2009, and reported by [PAR 13] and [RID 11], reveals that, out of a sample of 36,981 terms proposed by the Internet users of the steve.museum project, 86% of them are different from the controlled vocabularies and thesauruses used by professionals, 70.2% partially correspondent to the terms of the *Art and Architecture Thesaurus* in particular and 88.2% of its 36,981 terms were considered to be useful by these same professionals. An experiment at the Rijksmuseum in Amsterdam [OOS 14a], which compared the work of corrections, translations and identifications of flower species between experts and crowdworkers, also realized the relevance of using paid crowdsourcing for this type of task. A comparative study [ROR 10] of 4,441 tags on 1,000 Flickr photos compared to 3,709 descriptors on 996 photos of the photographic archive at the University of St. Andrew Library declared the complementarity of the professional approach and the folksonomic approach. This study encourages professionals to be inspired by the richness of the vocabulary used by Internet users in the construction of thesauruses. It confirms that a professional generally indexes a document on a subject with which he or she is unfamiliar, with the help of a thesaurus that is complex to use, and that he or she is trying to do it in order to allow a user to find information in a top-down, hierarchical approach. On the contrary, in the case of free indexing or *tagging*, the user very simply describes a document whose subject he or she usually knows well since his or her navigation generally has not led them to consult it by chance and he or she tags it more for their own interest in a bottom-up approach. If the terms that it uses are likely to be ambiguous, polysemous, synonymous and less precise, they nevertheless risk being richer and closest to the keywords entered as part of research. In addition to the simple production of data, Blasco *et al*. [BLA 13] has sought to compare the results between a team of developers organized in a traditional way and a team of self-organized developers among the crowd of developers within the framework of a competition with financial reward between the teams. The competitors coming from the crowd of developers had proposed more functioning and better-quality solutions.

Unlike the studies that support the benefits of using the work of amateurs, other studies are more nuanced. Thus, Bar-Ilan *et al*. [BAR 08] sought to compare free indexing resulting from amateur tagging with structured indexing carried out by professionals from thesauruses. 47 students in the information sciences participated in the experiment. They were divided into two groups: the first for free indexing and the second for structured indexing starting from fields to complete. The same images were given to them to be indexed. More detailed information was obtained by the second group that structured the information in the form of fields. This experiment, nevertheless, proves above all that the act of entering metadata into fields makes it possible to not forget the types of metadata. On the other hand, the qualities of the keywords that come from free indexing and structured indexing were not compared by the study.

Another study also appears to be relatively nuanced [SNO 081]. In this way, the quality of the annotations in natural language proposed by professionals and amateurs from the Amazon Mechanical Turk Marketplace was compared. According to the authors, the annotations obtained via amateurs are more numerous, but also more chaotic and less relevant than those produced by experts. Individually, the quality of the contributions is obviously better on the side of the experts. On the other hand, in comparing the work of four amateurs on average (two minimum and nine maximum), we obtain somewhat similar quality.

Finally, a study is clearly less enthusiastic regarding the quality of the data produced by amateurs. Oomen *et al.* [OOM 10] reflected upon the quality of the indexing produced for audiovisual documents at the Netherlands Institute for Sound and Vision via the game Waisda. It was found that only 5.8% of tags occurred in the institute's thesaurus and that only 23.6% of them are present in the Cornetto base of words in the Dutch language. In the same way, regarding the quality of the tags on Flickr: The Commons, a 2006 study by Guy & Tonkin reported by Earle estimates, based on a sample, found that only 40% of tags occurred in the dictionary *Open Source Aspell* [EAR 14].

In light of all of these studies whose results remain contradictory, it thus remains difficult to arrive at a definitive point of view on the subject.

3.5.3. *Reintegration of the data produced*

Depending on the quality of the data collected, it is decided whether or not to integrate the data produced by amateurs into the information system, catalog or digital library. Nevertheless, there are two philosophies on the subject. Some institutions simply want to engage the public in their collections or to improve their image as part of an institutional communication around a fashionable subject: these institutions have the tendency to not use the data produced by Internet users. Other institutions really need the help of Internet users. These therefore have more of a tendency to use the work which was generously offered to them.

In the report commissioned by OCLC [SMI 12], the writer relates that, following a study conducted by OCLC on 76 sites, it turned out that only half of the sites using crowdsourcing reused the tags produced by Internet users and that a little more than a third of respondents claim that they do not index the metadata produced by crowdsourcing. This result, surprising at first glance, clearly comes from the fact that crowdsourcing was not really intended as a means of outsourcing tasks or that the quality of the contributions remains disparaged by the profession.

Stiller [STI 14] estimates that the data produced by users is precious and must be valued in the same way as the data produced by institutions. He also notes that professionals who still remain hesitant to accept these contributions often remain separate from institutional content for fear of devaluation of their content or from fear of abuse on the part of users. According to the author, these misgivings are often arbitrary, especially if a work community exists that can monitor the content produced.

Survey questions (Section 8)	Yes (%)	No (%)
Are you concerned with how the content of your site is used or repurposed?	28	72
Have you incorporated metadata (including tagging) created by users into your own metadata and description workflow?	39	61
Do you incorporate other user-contributed content (e.g. photographs, documents) into your site?	44	56
Does your system index user-supplied metadata?	61	39
Do you perform any spell-checking of user content or carry out orthographic verification of user content or control of the tags submitted by users (e.g. differences in capitalization or spelling, singular vs. plural, etc.)?	19	81

Table 3.8. *Use of social metadata made by cultural institutions, according to the OCLC study [SMI 11]*

3.5.4. *The legal status of contributions: crowdsourcing and the semantic web*

An interesting publication from [DJU 13] relates the experience of the YEAH! crowdsourcing project applied to the field of archives and the use of technologies of the semantic web and *Linked Open Data*. The YEAH! project is funded by the Swedish Governmental Agency for Innovation Systems (VINNOVA), in partnership with NordForsk, the Icelandic Centre for Research (RANNIS) and the Estonian Ministry for Economic Affairs and Communication. In this publication, the authors mention in particular that it cannot ethically be forbidden to reuse contributions produced for free by volunteers or bind them with restrictive licenses. Logically, data produced though volunteer work must be able to be, freely and without restrictions, reused by everyone on the web including by other information systems, via *Linked Open Data* technologies. This possible reuse is moreover another argument for projects to encourage contributions.

In every case, at the time when the volunteers create an account on the site, it is necessary for them to approve a contract stating what the status of the data that they produce will be and its distribution license.

3.6. The evaluation of crowdsourcing projects

According to the study conducted by OCLC [SMI 11], 91% (30 responses) of the respondents consider their crowdsourcing project a success. This same study identifies the criteria for success as they are perceived by the projects' managers.

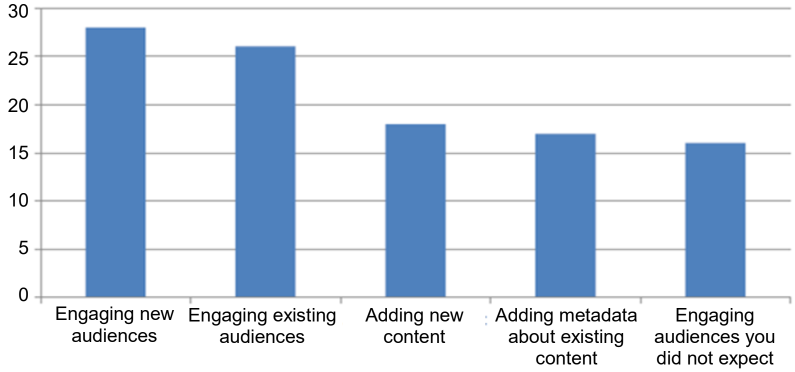
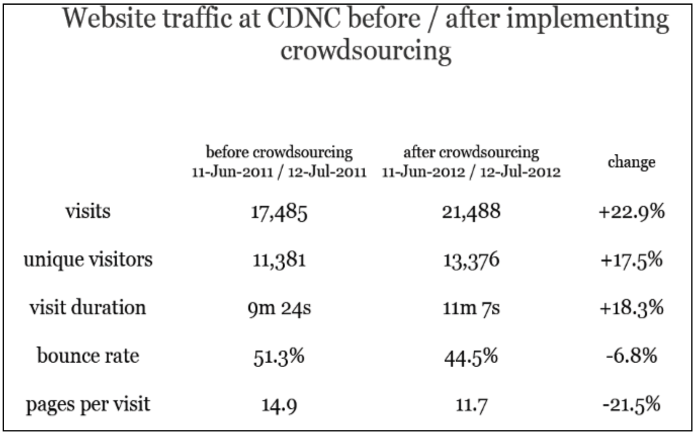


Figure 3.12. *The criteria for success, from [SMI 11]*

Above, we find the fact of engaging new and existing audiences, ahead of success in obtaining new content or adding metadata to already-existing content. It seems therefore that the philosophy that consists of launching crowdsourcing projects in order to change relation with the public and improve the institution’s image dominate to the detriment of the real usefulness of work of volunteers and real purposes for the projects. It therefore appears necessary to evaluate projects not only qualitatively but also quantitatively, an exercise to which institutions unfortunately do not always lend themselves. In doing this, tools such as Google Analytics, surveys and interviews are indispensable [BIR 12].

Crowdsourcing seems to have an impact on the web traffic of digital libraries. Nicole Saylor, head of the digital library at the University of Iowa, has reported that because of crowdsourcing, on June 9, 2011, the digital library went from 1,000 HITs maximum per day to more than 70,000 HITs[6].

Beyond this simple example, according to the OCLC survey [SMI 11], the number of unique visitors per month declared by cultural institutions is located as shown in Figure 3.13.



Website traffic at CDNC before / after implementing crowdsourcing

	before crowdsourcing 11-Jun-2011 / 12-Jul-2011	after crowdsourcing 11-Jun-2012 / 12-Jul-2012	change
visits	17,485	21,488	+22.9%
unique visitors	11,381	13,376	+17.5%
visit duration	9m 24s	11m 7s	+18.3%
bounce rate	51.3%	44.5%	-6.8%
pages per visit	14.9	11.7	-21.5%

Table 3.9. *Statistics before and after crowdsourcing for the California Digital Newspaper Collection, from [GEI 12]*

[6] The use of HIT in order to measure web traffic on a site does not seem to us to be the most relevant to the extent that the number of objects per web page can bias the results and give the impression that a web page consulted infrequently ends up generating more web traffic than a page that is consulted much more often, but contains fewer documents to display. Nevertheless, if we compare the same website at two different times, apart from large changes in content on this site, we can effectively conclude that there was a strong increase in its Web traffic as Nicole Saylor does.

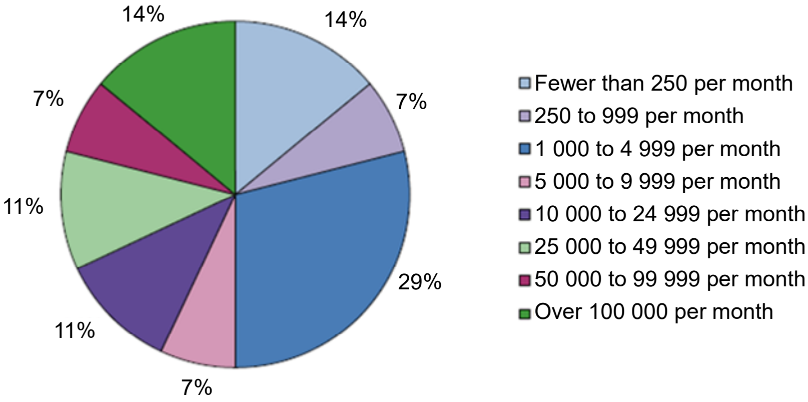


Figure 3.13. *Number of unique visitors per month for crowdsourcing projects, from [SMI 11]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

And, according to this same survey, the number of contributing visitors declared by the cultural institutions is shown in Figure 3.14.

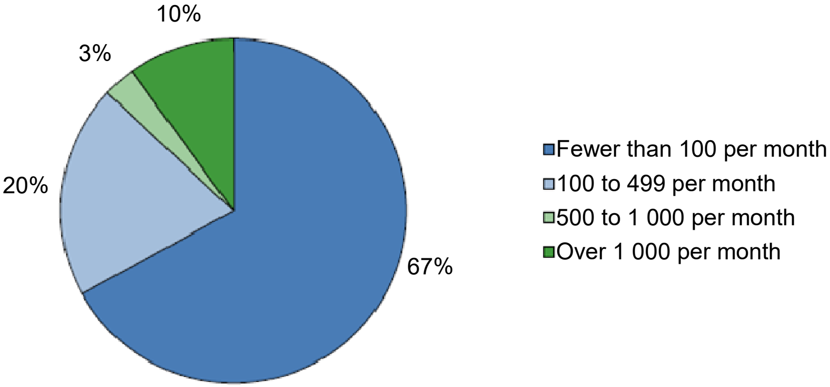


Figure 3.14. *Number of contributors per month for cultural institutions, from [SMI 11]. For a color version of the figure, see www.iste.co.uk/andro/libraries.zip*

3.6.1. *Factors in success and failure*

Taking inspiration from Sharma [SHA 10], Brokfeld's [BRO 12] thesis and McKinley's [MCK 15] studies, which are interested in the question of the factors in the success of crowdsourcing projects and without returning to all of the elements already mentioned, we can list the following keys to success:

– a vision, goal, purpose, or clearly defined challenge, expressed on the home page and easily understood by potential contributors. To return to the metaphor employed previously, the stonemason must know what his efforts will be used for; he must know that they will be used to build a cathedral. He must not be used as a simple means who does not have to know the context and the purpose of the work which is asked of him. He must be convinced that his work is indispensable and cannot be automated. The work instructions must be clear and effective;

– effective communication in order to recruit volunteers by laying out the various reasons to contribute and by playing on different types of motivation. The fact that they serve the public interest is an advantage for public cultural institutions. In order to be credible and appear sincere in the eyes of Internet users, the project must show that it has well-known sponsors and the project's team, its results and the number of contributors. The data produced by Internet users must be freely accessible and reuseable by the public;

– available and motivated human capital. As we have seen in the chapter about motivations, motivation can be stimulated by a table of the results, by a diagram of the progress toward the realization of the project, by ranking the biggest contributors, by acknowledgements and by rewards, etc.;

– an infrastructure likely to receive the work provided with an intuitive, ergonomic and reliable work environment. Users must be able to log in directly from their Facebook or Google+ accounts and easily invite the members of their social networks to participate;

– easy, fun, interesting and educational activities. New content to work on is regularly added. The new contributors must be able to train in a "sandbox". They must then be able to choose what they will work on according to their areas of interest, their levels of expertise and the time that they can dedicate to the project;

– trust and the social bond. Users must be able to interact and form a community. Volunteers must be given decision-making powers, have trust in them and earn their trust. The majority of successful projects that succeeded considered Internet users as partners in the project and not as a source of free labor. They truly used and reintegrated the data produced;

– internal change management and a favorable outside environment (economic and social situation).

Conversely, the projects that fail generally require tasks that are too complex, knowledge that is too specialized, following difficult tutorials or preliminary training, are too vague in their goals, not providing enough return to contributors and not showing them how data produced are used [RID 13].

More generally, the factors in a crowdsourcing project's failure are overall the same as all outsourcing with the dilution of responsibility. With outsourcing, the consequences of decisions are felt less, which can cause a loss of attention and control and make things take longer. The principal risk of crowdsourcing would therefore be that a business making use of this type of outsourcing overestimates its capacities to ensure management and that it ends up becoming completely unmanageable. It could also overestimate the profits that it could make from crowdsourcing.

Using crowds can also generate strong tensions among the internal staff. In fact, the most skeptical have a tendency to believe that allowing the profane to index is opening up cultural heritage to fan mail, false memories to the manipulation of memories or even incivility. Losing control of cultural heritage is also a source of anxiety, especially when it comes to their reuse in unethical contexts or their commercial reuse. Being swamped by comments to which they must respond can also be a subject of concern for librarians. Thus, according to a Swiss survey of libraries [EST 14], 72% of those surveyed estimated that implementing crowdsourcing will require considerable time for monitoring and that this time is difficult to estimate (70%) for unpredictable results (61%). Finally, crowdsourcing can also be considered a challenge to the indexing work of professionals, if the work is done for free by volunteers.

3.6.2. *Quantitative evaluation of crowdsourcing projects and their costs*

Just as it is essential for an organization to know if the outsourcing of an activity is less expensive and more productive than if it is maintained internally, outsourcing it to crowds of Internet users must be compared with its outsourcing to a more traditional subcontractor with its automatization. We will be comparing, in particular, the costs of crowdsourcing in communication, project management and community management, quality control, reward or payment with the costs as they exist when they are handled internally. However, according to [LEB 15], crowdsourcing would be "particularly competitive in terms of costs" and "in every case [...] less expensive than another form of outsourcing".

Correcting OCR or transcribing manuscripts	Number of words corrected or transcribed per month	Number of lines corrected or transcribed per month
reCAPTCHA	86 million words validated per day, which equals 2.58 billion words validated per month	Unknown
TROVE	Unknown	2,724,671 lines corrected (over the month of April 2015)
Digitalkoot	6,61,659 words corrected in 1 year, in February 2012, which is 538,472 words corrected per month	Unknown
Transcribe Bentham	From January 28, 2012 to November 2, 2012, on average 51 manuscripts (25,500 words) per week or 102,000 words per month	Unknown
California Digital Newspaper Collection (CDNC)	Unknown	578,000 lines in 2012, or 48,167 lines per month

Table 3.10. *Indicators of quantitative analysis of OCR correction or transcription projects*

Indexing content	Number of tags added per month
TROVE	68,167 tags (in the month of April 2015)
Flickr: The Commons	2 million tags in five years for 33,333 tags per month
steve.museum	468,120 tags between March 2007 and the end of 2010 which is 468,120/46 = 10,175, which is five tags per month

Table 3.11. *Indicators of quantitative analysis of content indexing projects*

Finance digitization of collections by Internet users through a digitization on demand crowdfunding service	Number of books digitized per month	Money collected per month
eBooks on Demand in 2011 (27 European libraries)	1,781 orders over the year 2011 = 148,50 books per month	€75,512,30 over the year 2011 = €6,542,69 per month
Livre à la carte, Phénix éditions	One hundred orders per month beginning in 2011 at the Bibliothèque Municipale de Troyes	Unknown
Adopt a book on Gallica	7,64 digitizations per month	Unknown
Numalire in 2014 (8 Parisian libraries)	36 books over 8 months = 4.50 books per month	Unknown

Table 3.12. *Indicators of quantitative analysis of digitization on demand projects*

Here are the indicators of quantitative analysis that we provide for each goal that an institution which is launching a crowdsourcing project can provide, with data for the projects when we were able to collect them as part of the overview of projects.

In addition to the data that we have collected in the course of the overview of projects, it would be interesting to produce the indicator presented in Table 3.13.

Convert the passive consumers of the digital library into active producers	Relationship between the number of visitors and the number of contributions
Increase the visitors to the digital library	Number of visits per month and per document
Carry out institutional communication	Number of articles mentioning the project in the local, national or international press, scientific journals, blogs, social networks, etc.
Reduce costs	Evaluate based on a given hourly wage, the money that the crowd has provided to the project in working time

Table 3.13. *Other indicators of evolution*

Citizen science is time- and energy-expensive in order to manage volunteers, websites and databases. However, the question of development costs for crowdsourcing projects is only very rarely mentioned in the literature. As Sagot *et al.* [SAG 11] also emphasize, development of platforms and quality control is only rarely evaluated in studies. Crowdsourcing applied to cultural heritage remains in an experimental phase and the initiatives are not always profitable [MCK 15].

One of the rare publications to mention this aspect is that of Causer *et al.* [CAU 12b]. This study provides precise indications of the costs of developing the Transcribe Bentham project (£262,673, two full-time research associates, a curator, a consultant). He also estimates that the two full-time research associates would have succeeded in transcribing about 5,000 manuscripts working full-time over 12 months, which is more than what Internet users produced. He thus considers that, in 2012, the Transcribe Bentham project made it possible to avoid spending the equivalent of hiring a full-time editor for 6 months. However, as this same author reports, the editors of the Papers of Abraham Lincoln ended up estimating that they spent more time correcting the corrections than doing the transcriptions themselves.

For all projects, it would be very interesting to compare the costs (labor, development, hosting, project management, communication, etc.) necessary for their development with the "benefits" that they have generated by converting the working time provided into value.

In the chapter dedicated to reCAPTCHA, we have estimated the amount that Google has not had to spend each year because of the involuntary work of Internet users at 146 million euros per year. In comparing these calculations with those of [GEI 12, IPE 11, ZAR 14], we estimate what the following projects did not have to spend on human OCR correction (if they had used service providers). The costs are shown in Table 3.14.

These unspent costs deserve, however, to be compared to the costs which were agreed upon in development, platform administration, communication, community management and reintegration of the data produced by volunteers.

Project	Type of crowdsourcing	Amount not spent
California Digital Newspaper Collection (end of 2011–2014)	Explicit crowdsourcing	$53,130 cumulative in June 2014
TROVE (August 2008–2014)	Explicit crowdsourcing	$2,580,926 cumulative in May 2014
Digitalkoot (February 2011)	Gamification	Between €31,000 and €55,000 cumulative in October 2012
Google Books and reCAPTCHA (2008)	Implicit crowdsourcing	146 million euros per year (at the 2008 rate)

Table 3.14. *Calculation of what OCR correction would have cost without use of crowdsourcing for several representative projects, from [AND 15]*

Regarding the digitization through crowdsourcing project Numalire, we have found that libraries dedicated 207 h of somewhat thankless work (formal descriptions of documents in order to produce estimates) to the project during the 8 months of the experiment and that we can estimate that this would have cost them €6,210 in working time for only 36 books whose digitization have been digitized, for an average of €172 per book. In this case, the benefits of the project thus have difficulty offsetting its costs.

3.7. Change management

Any change in working methods has a tendency to generate resistance and an almost natural desire to keep the older method of working. For libraries, crowdsourcing is a major change since the tasks which were previously performed by professionals will be able to be done by amateurs whose activities will change.

Resistance to change can take the following forms:

– inertia, procrastination, latent resistance;

– argued resistance;

– action against the change;

– sabotage or excess of enthusiasm;

– in libraries these types of reluctance are to be expected:

- the fear of having a workload that is too large to ensure control of the quality of the data produced by Internet users or as part of digitization on demand, to communicate the documents to the service provider and carry out observations of the states before and after digitization;

- ideological and cultural hostility toward private citizens, amateurs, public service delegations (refusal to subcontract public services by private businesses) and sponsorship (not very practical in a library);

- incredulity: no one will agree to work for us for free or participate in financing the digitization of books, particularly since the digitization of a book is still very expensive.

These reactions of resistance to change are natural since every individual is afraid of leaving behind what he or she knows for what they still do not know, but this can be very variable from one individual to another. A change in the Integrated library system (ILS) could be seen as a small change for a computer scientist, while it would be experienced as very difficult for a librarian forced to change his or her habits and to learn to work with a new interface.

The duration and the intensity of resistance to change decrease with better understanding of the project and participation in its implementation. The implementation of this change must be accompanied by solid change management. Otherwise, it will be a failure. This underestimation is moreover one of the major reasons that some crowdsourcing projects have failed.

Thus, we have an interest in having a typology of individuals and groups of individuals involved in a project according to the following criteria:

– the degree of support or opposition to the project;

– the power to influence the project (Who retains the knowledge? The skills? Who decides? Who controls the rules? etc.)

As Danièle Imbault emphasizes, the vast majority of individuals will generally be classified as reticence-passivity or as hesitation. Some are eager or unconditional pioneers. Others are hostile opponents or undecided. It is useless to waste energy communicating to the latter except in order to obtain the list of arguments against the project. On the other hand, it is necessary to concentrate efforts on the hesitant and the passive who have influence over others (key employees, opinion leaders, employee representatives) in order to convince them, to mobilize them around the project and to earn loyalty and lead the passive through influence. Otherwise, it is the opponents who will take charge. It is therefore better to seem more organized than they are and limit their power of influence by mobilizing the hesitant and passive.

To do this, several resources can be used: communication about the means, goals and purposes of the change, training, involvement in the change, support of independent experts or opinion leaders, recognition of the involvement of agents and the example of success of certain people (internal communication, party, etc.) and the taking into account of the point of view of volunteers, within the scope of meetings or in the form of forums where they can remain anonymous, regular trips to the "ground" in order to understand the agents and "evangelize", individual interviews, recruitment of people favorable to the project, evaluation regarding the project. As François Dupuy suggests in *La sociologie du changement* (in English: The Sociology of Change), "when it comes to change, the best is the enemy of the good and finicky perfectionism is a powerful factor in inaction".

In libraries, change is made difficult by the fact that "the executives have an administrative and management culture unfavorable to initiatives and innovation" [DEL 14]. Libraries also possess advantages such as the feeling, specific to public service, of working for the public interest. Nevertheless, libraries are used to change, they have already experienced the development of open access, electronic periodicals and digitization, they thus benefit from significant experience in change management.

Change takes some time, which can be regarded as a mourning period during which individuals permanently renounce the old mode of operation by going through the following phases:

– refusal to understand and denial, especially if the individual is satisfied by the current mode of operation and that it needs to be changed. They will then ignore the information;

– resistance to change, anger, shock, negotiation: it nevertheless makes it possible to better be aware of change and therefore its usefulness. Above all it makes it possible to explain the arguments against change which are likely, depending on their relevance, to reinforce them;

– preintention: from lack of information, from fear, lack of trust or interest, the person does not envisage immediate change in habits;

– intention: the individual hesitates and contemplates changing behavior soon;

– preparation: the individual decides to change and seeks to get training;

– action: the individual resigns herself and changes her habits;

– maintenance: with some temptation to move backwards;

– acceptance and resolution: the individual is satisfied with the change and the new mode of operation.

Consequently, change management will involve three phases:

– diagnostic;

– support;

– consolidation.

The ideal leader is an integrator who has both a strong interest in the results and for his or her collaborators. One of the main functions of managers and directors of libraries should be to have a vision of the library's future, to set the course, to give meaning to the changes, to explain the projects, to persuade people of the merit of the goals, purposes, to encourage and to motivate the participation of its teams. Management must know how to give them the feeling of being useful, benefitting from change, and knowing their responsibilities and room to maneuver.

As Lebraty and Lobre [LEB 15] report, businesses have many difficulties making use of crowdsourcing for the tasks that they previously carried out internally. On the other hand, it is much easier for them to implement a crowdsourcing approach as part of the deployment of a new activity.

This chapter has established an overview of crowdsourcing in digital libraries. In it we have proposed a taxonomy of projects and mentioned communication for recruiting, the types of motivations of the contributors, their sociology, the quality of the contributions, their reintegration, their legal status, the evaluation of projects and change management.

To conclude this chapter, and taking inspiration from [TWE 12], here are the main steps of a project that summarizes the points which were previously mentioned:

– before beginning, defining the purpose and choose the type of crowdsourcing to use. Therefore, depending on the type of crowdsourcing (explicit, gamification or implicit, volunteer or paid) and of the type of motivations of contributors (intrinsic or extrinsic), we will mobilize a particular type of contributor and obtain a result that is more or less effective. For example, if the work required consists of microtasks which provide absolutely nothing in terms of personal development, it might be difficult to find volunteers and might be preferable to resort to paid crowdsourcing;

– in the beginning of the project, recruit the project and human resources team, define its goals, identify the sources of the budget, identify the types of participants;

– in the development phase, devise the architecture, identify the technical, data and storage prerequisites, improve functioning iteratively;

– in the industrial phase, conduct a communication campaign around the project, integrate the data and provide a quick return to the contributors;

– in the analysis phase, reintegrate the data produced, make it reusable and evaluate the project.

Conclusion

We have studied the definition, origin, conceptual policy, economic and managerial philosophy of crowdsourcing in libraries, and then created an overview of existing projects to analyze this movement on the level of the taxonomy of projects, the motivation of Internet users, *community management*, the quality of the data produced, the evaluation of projects and change management. Over the course of this work, we have been led to develop our own concepts as contributions to the knowledge of crowdsourcing in libraries with original contributions on the subject of the historical origins of crowdsourcing, its taxonomy, the law of value and personal analyses on the subject of the difficulties of implementing it in libraries and the diametrically opposing conceptions leading to it.

Our starting hypothesis was that crowdsourcing and crowdfunding in particular could be profitable to the libraries that use them. Our analysis of the literature has allowed us to identify numerous advantages as much at the level of costs as the level of results from both a quantitative and qualitative point of view. It has also allowed us to identify disadvantages mostly around the question of the quality of the data produced, of a sometimes low return on investment compared to the cost, and more generally, around the question of the replacement of professionals by volunteers or underpaid workers.

From our point of view, organizations that set up the first participative approaches can stand out and gain a competitive advantage, but when the practice becomes widespread, it is very likely that gamification, rewards and payment are becoming the only means of gaining the participation of Internet users. However, as Holley [HOL 10c] suggests, libraries have much interest in mutualizing crowdsourcing instead of undertaking individual and competitive

Digital Libraries and Crowdsourcing, First Edition. Mathieu Andro.

approaches on their own. In this way, they would gain more volunteers by offering more content to correct.

Libraries have already lost their monopoly as unavoidable intermediaries between information and the public. Library curators could therefore have trouble with seeing their skills devalued and placed on the same level as those of the general public and also have the impression that anyone is being allowed to do anything, that the producer/consumer, professional/amateur hierarchy is being erased, and that the paid labor of the librarian is being replaced by the free or “Uberized” work of volunteers. Nevertheless, this “Uberization” of libraries seems inescapable and could also be very promising.

Bibliography

[ACA 11] ACAR O.A., VAN DEN ENDEN J., "Motivation, reward size and contribution in idea crowdsourcing", *Idea Connection, Build on the genius of the others*, available at: https://www.ideaconnection.com, p. 29, January 2011.

[ALA 12] ALAM S.L., CAMPBELL J., "Crowdsourcing motivations in a not-for-profit GLAM context: the Australian newspapers digitisation program", *Proceedings of the 23rd Australasian Conference on Information Systems 2012*, pp. 1–11, ACIS, Geelong (Vic), Australia, 2012.

[ALA 13a] ALAM S.L., CAMPBELL J., "Dynamic Changes in Organizational Motivations to Crowdsourcing for GLAMs", *34th International Conference on Information Systems*, Milan, Italy, p. 17, 2013.

[ALA 13b] ALAM S.L., CAMPBELL J., "A conceptual framework of influences on a non-profit GLAM crowdsourcing initiative: A socio-technical", *24th Australasian Conference on Information Systems Socio-technical Model of Crowdsourcing Influences*, Melbourne, Australia, 2013.

[AND 10] ANDERSON R., "The Espresso Book Machine: The Marriott Library Experience", *Serials*, vol. 23, no. 1, pp. 39–42, 2010.

[AND 12] ANDRO M., ASSELIN E., MAISONNEUVE M., "Digital libraries : comparison of 10 software", *Library Collections, Acquisitions, and Technical Services*, vol. 36, no. 3–4, pp. 79–83, 2012.

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

Digital Libraries and Crowdsourcing, First Edition. Mathieu Andro.

[AND 14b] ANDRO M., RIVIERE 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*, p. 9, 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*, p. 8, 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*, p. 7, 2015.

[AND 16] ANDRO M., Crowdsourcing et bibliothèques numériques : expérimentations autour de Numalire, projet de numérisation à la demande par crowdfunding, Doctoral thesis, Paris 8 University, 2016.

[AND 17] ANDRO M., SALEH I., “Digital Libraries and Crowdsourcing : A Review”, in SZONIECKY S., BOUHAÏ N. (eds), *Collective Intelligence and Digital Archives*, ISTE Ltd, London and John Wiley & Sons, New York, 2017.

[ARL 11] ARLITSCH K., “The Espresso Book Machine: a change agent for libraries”, *Library Hi Tech*, vol. 29, no. 1, pp. 62–72, 2011.

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

[BAI 12] BAINBRIDGE D., TWIDALE M.B., “Interactive context-aware user-driven metadata correction in digital libraries”, *International Journal on Digital Libraries*, no. 13, pp. 17–32, 2012.

[BAR 00a] BARBROOK R., *L’économie du don high-tech, Libres enfants du savoir numérique*, Editions de l’Eclat, Paris, 2000.

[BAR 00b] BARBROOK R., “Cyber-communism : how the Americans are superseding capitalism in cyberspace”, *Science As Culture*, 2000.

[BAR 00c] BARBROOK R., CAMERON A., *The Californian Ideology: Revised SaC Version*, Borsook, 2000.

[BAR 08] BAR-ILAN J., SHOHAM S., IDAN A. *et al.*, “Structured *versus* unstructured tagging : A case study”, *Online Information Review*, vol. 32, no. 5, pp. 635–647, 2008.

[BAU 10] BAUER A., “Sciences participatives et biodiversité : implication du public, portée éducative et pratiques pédagogiques associées”, *Les livrets de l’Ifrée*, no. 2, p. 107, 2010.

[BAU 15] BAUWENS M., *Sauver le monde : vers une économie post-capitaliste avec le peer-to-peer*, Les liens qui libèrent, Paris, 2015.

[BEN 14] BENYAYER L.D., “Open Models : les *business models* de l’économie ouverte”, available at: http://www.openmodels.fr/en, p. 226, 2014.

[BEU 14] BEUTH HOCHSCHULE FÜR TECHNIK, “Charting diversity: working together towards diversity in Wikipedia”, *Wikimedia Deutschland*, available at: https://upload.wikimedia.org/wikipedia/commons/5/57/Charting_Diversity.pdf, p. 21, 2014.

[BIE 15] BIELLA D., SACHER D., WEYERS B. *et al.*, “Crowdsourcing and Knowledge Co-creation in Virtual Museums”, *Proc. International Conference on Collaboration and Technology (CRIWG)*, p. 18, 2015.

[BIR 12] BIRCHALL D., HENSON M., BURCH A. *et al.*, “Levelling Up: Towards Best Practice in Evaluating Museum Games”, available at: http://www.museumsandtheweb.com/mw2012, p. 11, 2012.

[BLA 13] BLASCO A., BOUDREAU K.J., LAKHANI K.R. *et al.*, “Do Crowds have the Wisdom to Self-Organize?”, available at: http://scholar.harvard.edu/ablasco, 2013.

[BLU 16] BLUMMER B., “Opportunities for Libraries with print-on-demand Publishing”, *Journal of Access Services 3*, no. 2, pp. 41–54, 2016.

[BOE 12] BŒUF G., ALLAIN Y.-M., BOUVIER M., L’apport des sciences participatives dans la connaissance de la biodiversité, Report submitted to the French Ecology Minister, Paris, 2012.

[BON 09] BONNEY R., BALLARD H., JORDAN R. *et al.*, Public Participation in Scientific Research : Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report, Center for Advancement of Informal Science Education (CAISE), Washington, DC, 2009.

[BRA 10] BRABHAM D.C., “Moving the crowd at threadless”, *Information, Communication and Society*, vol. 13, no. 8, pp. 1122–1145, 2010.

[BRA 12] BRABHAM D.C., “The myth of amateur crowds : A critical discourse analysis of crowdsourcing coverage”, *Information, Communication & Society 15*, no. 3, pp. 394–410, 2012.

[BRO 12] BROKFELD J., Die digitale Edition der “preußischen Zeitungsberichte”: Evaluation von Editionswerkzeugen zur nutzergenerierten Transkription handschriftlicher Quellen, Master’s Thesis, 2012.

[BRU 12] BRUMEN M., BLATNIK A., Recent developments and results of the european library project Ebooks on Demand (EOD), National and University Library, Slovenia, vol. 21, pp. 87–93, 2012.

[CAR 10] CARDON D., *La démocratie Internet : promesses et limites*, Le Seuil, Paris, 2010.

[CAR 13] CARTELLETI L., GIANNACHI G., PRICE D. *et al.*, “Digital Humanities and Crowdsourcing: An Exploration”, *MW2013 – Museums and the Web 2013 : The annual conference of Museums and the Web*, p. 17–20, Portland, United States, April 2013.

[CAR 16] CARDON D., CASSILLI A., “Qu’est-ce que le *digital labor* ? Les enjeux de la production de valeur sur Internet et la qualification des usages numériques ordinaires comme travail”, *Etudes et controverses*, INA, Paris, 2016.

[CAS 11] CASEMAJOR LOUSTAU N., “La contribution triviale des amateurs sur le Web : quelle efficacité documentaire ?”, *Etudes de communication*, no. 1, pp. 39–52, 2011.

[CAU 12a] CAUSER T., WALLACE V., “Building A Volunteer Community : Results and Findings from Transcribe Bentham”, *Digital Humanities Quarterly*, vol. 6, no. 2, p. 26, 2012.

[CAU 12b] CAUSER T., TONRA J., WALLACE V., “Transcription maximized ; expense minimized ? Crowdsourcing and editing The Collected Works of Jeremy Bentham”, *Literary and Linguistic Computing*, vol. 27, no. 2, pp. 119–137, 2012.

[CHA 10] CHAMBERLAIN E., “Digitisation-on-Demand in Academic Research Libraries”, available at: https://www.repository.cam.ac.uk, p. 62, 2010.

[CHA 12] CHAMBERLAIN E., “Investigating Faster Techniques for Digitization and Print-on-Demand”, *New Review of Academic Librarianship*, vol. 18, no. 1, pp. 57–71, 2012.

[CHA 13] CHAUDIRON S., “Ordres et désordres numériques”, in GERMAIN M., PERALES C., BUFFARD P. *et al.* (eds), *Les organisations du XXI^{e} siècle*, Documentaliste-Sciences de l’Information, no. 50, pp. 38–47, 2013.

[CHA 15] CHARDONNENS A., Collections iconographiques numérisées et crowdsourcing : possibilités et limites de la co-création de métadonnées par le grand public au travers de trois études de cas, Master’s Thesis, ULB, 2015.

[CHR 11] CHRONS O., SUNDELL S., “Digitalkoot: Making Old Archives Accessible Using Crowdsourcing”, *HCOMP 2011: 3rd Human Computation Workshop*, San Francisco, United States, 2011.

[CHU 06] CHUN S., CHERRY R., HIWILLER D. *et al.*, "Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums", in TRANT J., BEARMAN D. (eds), *Museums and the Web 2006 : Proceedings*, Archives & Museum Informatics, Toronto, 2006.

[CIT 14] CITTON Y., *Pour une écologie de l'attention*, Le Seuil, Paris, 2014.

[COH 15] COHEN D., *Le Monde est clos et le désir infini*, Albin Michel, Paris, 2015.

[COL 13] COLQUHOUN B., "Making Sense of Historic Photographic Collections on Flickr The Commons : Institutional and User Perspectives", in PROCTOR N., CHERRY R. (eds), *Museums and the Web 2013*, *Annual Conference*, Silver Spring, United States, 2013.

[CON 09] CONTEH A., TZADOK A., "User collaboration in mass digitisation of textual materials", *Proceedings Cultural Heritage Online. Empowering users: an active role for user communities*, Liber Library, available at: http://libereurope.eu, p. 7, 2009.

[CRO 13] CROWSTON K., PRESTOPNIK N.R., "Motivation and data quality in a citizen science game : A design science evaluation", *46th Hawaii International Conference on System Sciences* (HICSS-46), p. 10, 2013.

[DAE 09] DAELE A., "Les communautés de pratique", in BARBIER J.-M., BOURGEOIS E., CHAPELLE G. *et al.* (eds), *Encyclopédie de la formation*, PUF, Paris, 2009.

[DAU 14] DAUDEY E., HOIBIAN S., "La société collaborative : mythe et réalité", *Cahier de recherche*, no. 313, p. 65, CREDOC, 2014.

[DAW 11] DAWSON R., BYNGHALL S., "The rise of crowdsourcing", *Getting Results From Crowds: The definitive guide to using crowdsourcing to grow your business*, pp. 9-12, Advanced Human Technologies Inc., San Francisco, 2011.

[DEB 12] DE BOER V., HILDEBRAND M., AROYO L. *et al.*, "Nichesourcing : Harnessing the Power of Crowds of Experts. Knowledge Engineering and Knowledge Management", *Lecture Notes in Computer Science*, no. 7603, pp. 16–20, 2012.

[DEL 14] DELAINE V., L'accompagnement du changement en bibliothèques : une approche managériale, Thesis, ENSSIB, Villeurbanne, 2014.

[DEN 13] DENG X.N., JOSHI K.D., "Is crowdsourcing a source of worker empowerment or exploitation? Understanding crowd workers' perceptions of crowdsourcing career", *34th International Conference on Information Systems*, p. 10, Milan, Italy, 2013.

[DEO 14] DEODATO J., "The patron as producer: libraries, web 2.0, and participatory culture", *Journal of Documentation*, vol. 70, no. 5, pp. 734–758, 2014.

[DET 11b] DETERDING S., SICART M., NACKE L. *et al.*, "Gamification : Using Game Design Elements in Non-Gaming Contexts", *Proceeding CHI '11 Extended Abstracts on Human Factors in Computing Systems*, Vancouver, Canada, pp. 2425–2428, 2011.

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

[DJU 13] DJUPDHAL M., ESKOR E., JONSSON O. *et al.*, "Review of Crowdsourcing Projects and how they relate to Linked Open Data", available at: http://www.itu.diva-portal.org, p. 15, February 2013.

[DOU 09] DOUGHERTY W.C., "Print on Demand : What Librarians should know", *The Journal of Academic Librarianship*, vol. 35, no. 2, pp. 184–186, 2009.

[DUN 12] DUNN S., HEDGES M., Crowd-Sourcing Scoping Study Engaging the Crowd with Humanities Research, Centre for e-Research, Department of Digital Humanities, King's College London, London, 2012.

[DUN 13] DUNN S., HEDGES M., "Crowd-sourcing as a component of humanities research infrastructures", *International Journal of Humanities and Arts Computing*, vol. 7, nos. 1–2, pp. 147–169, 2013.

[DWO 12] DWORAK M., "The Public as Collaborator : Towards Developing Crowdsourcing Models for Digital Research Initiatives", *B Sides: Journal of the University of Iowa School of Library and Information Science*, University of Iowa, 2012.

[EAR 14] EARLE E.F., Crowdsourcing metadata for library and museum collections using a taxonomy of Flickr user behavior, Thesis, Cornell University, 2014.

[EIC 12] EICKHOFF C., HARRIS C.G., DE VRIES A.P. *et al.*, "Quality through Flow and Immersion : Gamifying Crowdsourced Relevance Assessments", *SIGIR '12 Proceedings of the 35th International ACM SIGIR conference on Research and development in information retrieval*, pp. 871–880, 2012.

[ELL 14] ELLIS S., "A History of Collaboration, a Future in Crowdsourcing: Positive Impacts of Cooperation on British Librarianship", *Libri*, vol. 64, no. 1, pp. 1–10, 2014.

[EST 12] ESTELLES-AROLAS E., GONZALEZ-LADRON-DE-GUEVARA F., "Towards an integrated crowdsourcing definition", *Journal of Information Science*, p. 14, 2012.

[EST 14] ESTERMANN B., "Diffusion of Open Data and Crowdsourcing among Heritage Institutions: Results of a Pilot Survey in Switzerland", *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 9, no. 3, pp. 15–31, 2014.

[EST 15] ESTERMANN B., “Diffusion of Open Data and Crowdsourcing among Heritage Institutions : Based on data from Finland, Poland, Switzerland, and The Netherlands”, *EGPA 2015 Conference*, Toulouse, France, p. 27, 2015.

[EVE 13] EVELEIGH A., JENNET C., LYNN S. *et al.*, “I want to be a Captain! I want to be a captain!: Gamification in the Old Weather Citizen Science Project”, *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, Toronto, Canada, ACM Digital Library, pp. 79–82, 2013.

[FLA 12] FLANAGAN M., CARINI P., “How games can help us access and understand archival images”, *The American Archivist*, no. 75, pp. 514–537, 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.

[FUC 12] FUCHS C., “Dallas Smythe Today – The Audience Commodity, the Digital Labour Debate, Marxist Political Economy and Critical Theory. Prolegomena to a Digital Labour Theory of Value”, *triple C*, vol. 10, no. 2, pp. 692–740, 2012.

[GEI 11] GEITGEY T., “The University of Michigan Library Espresso Book Machine experience”, *Library Hi Tech*, vol. 29, no. 1, pp. 51–61, 2011.

[GEI 12] GEIGER B., ZARNDT F., “No tempest in my teapot : analysis of crowdsourced data and user experience at the California Digital Newspaper Collection”, *Bulletin des bibliothèques de France*, p. 70, 2012.

[GOO 11] GOOD B.M., SU A., “Games with a scientific purpose”, *Genome Biology*, vol. 12, no. 135, p. 3, 2011.

[GÖT 14] GÖTTL F., “Crowdsourcing with gamification”, *Advances in Embedded Interactive Systems Technical Report*, vol. 2, no. 3, pp. 15–19, 2014.

[GOU 14] GOUIL H., “La place de la coopération dans une conception eudémoniste du travail et de l’échange économique”, in HAMMOND KETILSON L., ROBICHAUD VILLETTAZ M.P. (eds), Le pouvoir d’innover des coopératives : textes choisis de l’appel international d’articles scientifiques, available at: https://www.researchgate.net, 2014.

[GRO 12] GROH F., “Gamification : State of the Art Definition and Utilization”, *Proceedings of the 4th Seminar on Research Trends in Media Informatics*, pp. 39–46, Ulm University, Germany, 2012.

[GRO 13] GROSDHOMME LULIN E., *La République participative*, Paradigmes et cætera, Paris, 2013.

[GST 09] GSTREIN S., MÜHLBERGER G., “User-driven content selection for digitization : the eBooks on Demand Network”, *Proceedings of International Conference on Cultural Heritage*, p. 6, 2009.

[GST 11] GSTREIN S., MÜHLBERGER G., “Producing eBooks on Demand : A European Library Network”, in PRICE K., HAVERGAL G. (eds), *E-books in Libraries: A practical guide*, Facet Publishing, London, 2011.

[HAG 13] HAGON P., *Trove Crowdsourcing Behaviour*, National Library of Australia, Canberra, 2013.

[HAM 14] HAMARI J., KOIVISTO J., SARSA H., “Does Gamification Work? A Literature Review of Empirical Studies on Gamification”, *Proceedings of the 47th Hawaii International Conference on System Sciences*, p. 10, 2014.

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

[HOL 09a] HOLLEY R., “A success story: Australian Newspaper Digitisation Program”, *Online Currents*, vol. 23, no. 6, pp. 283–295, 2009.

[HOL 09b] HOLLEY R., “Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers National Library of Australia”, available at: https://www.nla.gov.au, 2009.

[HOL 09c] HOLLEY R., “Crowdsourcing and social engagement: potential, power and freedom for libraries and users”, *Pacific Rim Digital Library Alliance Annual Meeting*, Auckland, New Zealand, p. 28, 2009.

[HOL 09d] HOLLEY R., “How good can it get ? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs”, *D-Lib Magazine*, vol. 15, nos. 3–4, 2009.

[HOL 10a] HOLLEY R., “Tagging Full Text Searchable Articles: An Overview of Social Tagging Activity in Historic Australian Newspapers August 2008 – August 2009”, *D-Lib Magazine*, vol. 16, nos. 1–2, 2010.

[HOL 10b] HOLLEY R., “Trove: Innovation in Access to Information in Australia”, *Ariadne*, no. 64, p. 9, 2010.

[HOL 10c] HOLLEY R., “Crowdsourcing : How and Why Should Libraries Do It?”, *D-Lib Magazine*, vol. 16, nos. 3–4, 2010.

[HOL 11] HOLLEY R., “Resource Sharing in Australia : Find and Get in Trove – Making “Getting” Better”, *D-Lib Magazine*, vol. 17, nos. 3–4, p. 14, 2011.

[HOU 16] HOULLIER F., MERILHOU-GOUDARD J.-B., ANDRO M. *et al.*, Les sciences participatives en France, Report to the French Minister of Higher Education, 2016.

[HUB 09] HUBERMAN B.A., ROMERO D.M., WU F., “Crowdsourcing, attention and productivity”, *Journal of Information Science*, vol. 35, no. 758, 2009.

[HUV 08] HUVILA I., "Participatory archive : towards decentralised curation, radical user orientation, and broader contextualisation of records management", *Archival Science*, no. 8, pp. 15–36, 2008.

[IPE 10a] IPEIROTIS P.G., "Analyzing the Amazon Mechanical Turk Marketplace", *XRDS*, vol. 17, no. 2, pp. 16–21, 2010.

[IPE 10b] IPEIROTIS P.G., Demographics of Mechanical Turk, New York University, 2010.

[IPE 11] IPEIROTIS P.G., PARITOSH P.K., "Managing Crowdsourced Human Computation", *20th International World Wide Web Conference WWW 2011*, p. 5, 2011.

[KAU 11] KAUFMANN N., SCHULZE T., VEIT D., "More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk", *Proceedings of the Seventeenth Americas Conference on Information Systems*, Detroit, United States, p. 11, 2011.

[KIT 13] KITTUR A., NICKERSON J.V., BERNSTEIN M.S. *et al.*, "The Future of Crowd Work", *CSCW '13*, p. 17, 2013.

[KLE 14] KLEKA P., ŁUPKOWSKI P., "Gamifying science: the issue of data validation", *Homo Ludens*, vol. 1, no. 6, 2014.

[KLO 14] KLOPP S., Numérisation et impression à la demande en bibliothèque: un panorama, Thesis, ENSSIB, Villeurbanne, 2014.

[LAK 13] LAKHANI K., "Using the crowd as an innovation partner", *Harvard Business Review*, p. 12, 2013.

[LAL 15] LALLEMENT M., *L'âge du faire : hacking, travail, anarchie*, Le Seuil, Paris, 2015.

[LAN 11] LANG A.S.I.D., RIO-ROSS J., "Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents", *Code4lib Journal*, no. 15, pp. 10–31, 2011.

[LEB 15] LEBRATY J.F., LOBRE K., *Crowdsourcing*, ISTE Ltd, London and John Wiley & Sons, New York, 2015.

[LEC 13] LE CROSNIER H., NEUBAUER C., STORUP B., "Sciences participatives ou ingénierie sociale : quand amateurs et chercheurs co-produisent les savoirs", *Hermès*, *La Revue*, no. 67, pp. 68–74, 2013.

[LED 06] LE DEUFF O., "Folksonomies Les usagers indexent le web", *Bulletin des Bibliothèques de France*, vol. 51, no. 4, pp. 66–70, 2006.

[LED 15] LE DEUFF O., "Les humanités digitales précèdent-elles le numérique ?", in *H2PTM 15*, ISTE Editions, London, 2015.

[LEV 14] LEVI A.S., "Memorializing Religion : Crowdsourcing, Minorities, and the Quest for Identity in Online Archives", *Advances in the Study of Information and Religion*, vol. 1, no. 9, p. 23, 2014.

[LEW 10] LEWIS D.W., "The User-Driven Purchase Giveaway Library", *Educause Review*, vol. 45, no. 5, pp. 10–11, 2010.

[LIÈ 14a] LIÈVRE P., LAROCHE N., "Retour sur la notion de communauté épistémique", *7e Colloque GeCSO LEST CNRS,* Aix-Marseille Université, Marseille, p. 25, 2014.

[LIE 14b] LIEW C.L., "Participatory Cultural Heritage : A Tale of Two Institutions' Use of Social Media", *D-Lib Magazine*, vol. 20, nos. 3–4, p. 17, 2014.

[LIG 12] LIGEON L., Crowd sourcing: labour struggles revisited?, Thesis, Utrecht University, The Netherlands, 2012.

[MAC 1837] MACHIAVEL N., *Œuvres complètes*, vol. 1, Auguste Desrez, Paris, 1837.

[MCC 12] MCCARTHY S., "Using gamification as an effective OCR crowdsourcing motivator", p. 19, 2012.

[MCK 12a] MCKINLEY D., "A Cognitive Walkthrough of the What's the Score at the Bodleian? Task interface to increase volunteer participation", available at: http://www.academia.edu, p. 14, 2012.

[MCK 12b] MCKINLEY D., "Optimizing crowdsourcing websites for volunteer participation. A case study: What's on the Menu? ", *New York Public Library, National Digital Forum conference*, Victoria, University of Wellington, New Zealand, p. 14, 2012.

[MCK 12c] MCKINLEY D., "Practical management strategies for crowdsourcing in libraries, archives and museums", available at: http://www.digitalglam.org, p. 13, 2012.

[MCK 13a] MCKINLEY D., Functionality and usability requirements for a crowdsourcing task interface that supports rich data collection and volunteer participation. A case study: The New Zealand Reading Experience Database, Thesis, University of Wellington, New Zealand, 2013.

[MCK 13b] MCKINLEY D., How effectively are crowdsourcing websites supporting volunteer participation and quality contribution?, presented at Hamilton City Library, Hamilton, New Zealand, 2013.

[MCK 13c] MCKINLEY D., "Why evaluation isn't a party at the end : Evaluating crowdsourcing websites", *National Digital Forum Conference*, Wellington, New Zealand, p. 9, November 2013.

[MCK 14] MCKINLEY D., "Accélérer la mutation numérique des entreprises : un gisement de croissance et de compétitivité pour la France", available at: http://mckinsey.com, p. 134, 2014.

[MCK 15] MCKINLEY D., "Heuristics to support the design and evaluation of websites for crowdsourcing the processing of cultural heritage assets", available at: http://www.nonprofitcrowd.org, 2015.

[MCL 68] MCLUHAN M., *Pour comprendre les médias*, Le Seuil, Paris, 1968.

[MCS 11] MCSHANE I., "Public libraries, digital literacy and participatory culture", *Discourse: Studies in the Cultural Politics of Education*, vol. 32, no. 3, pp. 383–397, 2011.

[MEA 52] MEADE J.E., "External Economies and Diseconomies in a Competitive Situation", *The Economic Journal*, vol. 62, no. 245, pp. 54–67, 1952.

[MEY 11] MEYER M., MOLYNEUX-HODGSON S., "Communautés épistémiques": une notion utile pour théoriser les collectifs en sciences?", *Terrains & travaux*, vol. 1, no. 18, pp. 141–154, 2011.

[MIL 11] MILLERAND F., HEATON L., PROULX S., "Emergence d'une communauté épistémique: création et partage du savoir botanique en réseau", in PROULX S., KLEIN A. (eds), *Connexions : communication numérique et lien social*, Presses universitaires de Namur, 2011.

[MOI 13a] 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, available at: http://www.enssib.fr, 2013.

[MOI 13b] MOIREZ P., STUTZMANN D., "Signaler les ressources numérisées : enrichissement, visibilité, dissémination", *Manuel de constitution de bibliothèques numériques*, Editions du Cercle de La Librairie, Paris, 2013.

[MOI 13c] MOIREZ P., "Bibliothèques, crowdsourcing, métadonnées sociales", *Bulletin des Bibliothèques de France*, vol. 5, pp. 32–36, 2013.

[MOY 11] MOYLE M., TONRA J., WALLACE V., "Manuscript transcription by crowdsourcing: Transcribe Bentham", *Liber Quarterly, the Journal of European Research Libraries*, vol. 20, nos. 3–4, 2011.

[MUH 09] MÜHLBERGER G., GSTREIN S., "eBooks on Demand (EOD): a European digitization service", *IFLA Journal*, vol. 35, no. 1, pp. 35–43, 2009.

[NEL 12] NELSON M.J., "Soviet and American Precursors to the Gamification of Work", *Proceedings of the 16th International Academic MindTrek Conference*, Tampere, Finland, pp. 23–26, 2012.

[NGU 12] Nguyen L.C., Partridge H.L., Edwards S.L., "Towards an understanding of the participatory library", *Library Hi Tech*, vol. 30, no. 2, pp. 335–346, 2012.

[OLL 13] Ollikainen M., On gamification, Thesis, University of Tampere, 2013.

[ONN 13] Onnee S., Renault S., "Le financement participatif : atouts, risques et conditions de succès", *Gestion*, no. 38, pp. 54–65, 2013.

[ONN 14] Onnee S., Renault S., "Crowdfunding : vers une compréhension du rôle joué par la foule", *Management & Avenir*, vol. 8, no. 74, pp. 117–133, 2014.

[OOM 10] Oomen J., Brinkerink M., Heijmans L. *et al.*, "Emerging Institutional Practices : Reflections on Crowdsourcing and Collaborative Storytelling", in Trant J., Bearman D. (eds), *Proceedings, Toronto : Archives & Museum Informatics*, Toronto, Canada, p. 9, 2010.

[OOM 11] Oomen J., Aroyo L., "Crowdsourcing in the cultural heritage domain : opportunities and challenges", *5th International Conference on Communities and Technologies*, Brisbane, Australia, June-July 2011.

[OOS 14a] Oosterman J., Bozzon A., Houben G.-J. *et al.*, "Crowd vs. Experts: Nichesourcing for Knowledge Intensive Tasks in Cultural Heritage", *WWW'14 Proceedings & Companion*, pp. 567–568, Seoul, South Korea, 2014.

[OOS 14b] Oosterman J., Nottamkandathy A., Dijkshoorn C. *et al.*, "Crowdsourcing Knowledge-Intensive Tasks in Cultural Heritage", in Menczer F., Hendler J., Dutton W.H. *et al.* (eds), *ACM Web Science Conference, WebSci '14*, Bloomington, United States, pp. 267–268, 2014.

[ORG 10] Organisciak P., Why Bother? Examining the Motivations of Users in Large-Scale Crowd-Powered Online Initiatives, Thesis, University of Alberta, 2010.

[OWE 13] Owens T., "Digital Cultural Heritage and the Crowd", *Curator: The Museum Journal*, no. 56, pp. 121–130, 2013.

[PAR 13] Paraschakis D., Crowdsourcing cultural heritage metadata through social media gaming, Thesis, Malmö University, 2013.

[PAR 14] Paraschakis D., Gustafsson Friberger M., "Playful crowdsourcing of archival metadata through social networks", *2014 ASE Bigdata/Socialcom/ Cybersecurity Conference*, Stanford University, United States, p. 9, 2014.

[PET 08] Petersen S.M., "Loser Generated Content : From Participation to Exploitation", *First Monday*, vol. 13, no. 3, 2008.

[PEU 12] PEUGEOT V., “Biens communs et numérique : l’alliance transformatrice”, in L. CALDERAN, P. LAURENT *et al.* (eds), *Le document numérique à l’heure du web*, *Sciences et techniques de l’information*, INRIA, 2012.

[PEU 15] PEUGEOT V., BEUSCART J.S., PHARABOD A.S. *et al.*, “Partager pour mieux consommer ? Enquête sur la consommation collaborative”, *Esprit*, no. 7, pp. 19–29, 2015.

[PIG 13] PIGNAL M., PEREZ E., “Numériser et promouvoir les collections d’histoire naturelle”, *Bulletin des Bibliothèques de France*, vol. 58, no. 5, pp. 27–31, 2013.

[POE 12] POETZ M.K., SCHREIER M., “The Value of Crowdsourcing : Can Users Really Compete with Professionals in Generating New Product Ideas?”, *Journal of Product Innovation Management*, vol. 29, no. 2, pp. 245–256, 2012.

[QUI 11] QUINN A.J., BEDERSON B.B., “Human Computation : A Survey and Taxonomy of a Growing Field”, *CHI ‘11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1403–1412, 2011.

[RAD 10] RADDICK J., BRACEY G., GAY P.L. *et al.*, “Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers”, *Astronomy Education Review*, vol. 9, no. 1, p. 18, 2010.

[RAD 14] RADICE S., Designing for participation within cultural heritage: Participatory practices and audience engagement in heritage experience processes, Politecnico di Milano, Design Department, Milan, 2014.

[REI 08] REIGHART R., OBERLANDER C., “Exploring the future of interlibrary loan : generalizing the experience of the University of Virginia, USA”, *Interlending & Document Supply*, vol. 36, no. 4, pp. 184–190, 2008.

[REN 14a] RENAULT S., “Crowdsourcing: La nébuleuse des frontières de l’organisation et du travail”, *RIMHE: Revue Interdisciplinaire Management, Homme(s) et Entreprise*, vol. 2, no. 11, pp. 23–40, 2014.

[REN 14b] RENAULT S., “Comment orchestrer la participation de la foule à une activité de crowdsourcing? La taxonomie des 4 C”, *Systèmes d’information & management*, no. 19, pp. 77–105, 2014.

[RID 11] RIDGE M., “Playing with Difficult Objects: Game Designs to Improve Museum Collections”, in J. TRANT, D. BEARMAN (eds), *Museums and the Web 2011: Proceedings*, Archives & Museum Informatics, Toronto, 2011.

[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.

[RIF 96] RIFKIN J., *La Fin du travail: ou comment l’Europe se substitue peu à peu à l’Amérique dans notre imaginaire*, La Découverte, Paris, 1996.

[RIF 14] RIFKIN J., *La nouvelle société du coût marginal zéro: l'internet des objets, l'émergence des communaux collaboratifs et l'éclipse du capitalisme*, Les liens qui libèrent, Paris, 2014.

[ROG 11] ROGSTADIUS J., KOSTAKOS V., KITTUR A. *et al.*, "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets", *Proceeding of the Fifth International AAAI conference on Weblogs and Social Media*, p. 9, 2011.

[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.

[ROS 10] ROSS J., IRANI L., SILBERMAN M.S. *et al.*, "Who are the Crowdworkers ? Shifting Demographics in Mechanical Turk", *Proceeding CHI EA '10 CHI '10 Extended Abstracts on Human Factors in Computing Systems*, pp. 2863–2872, 2010.

[ROT 16] ROTH Y., Comprendre la participation des internautes au crowdsourcing : une étude des antécédents de l'intention de participation à une plateforme créative, Thesis, Paris 1 University, 2016.

[ROU 10] ROUSE A.C., "A Preliminary Taxonomy of Crowdsourcing", *ACIS Proceedings*, Papier 76, p. 10, 2010.

[SAB 13] SABOU M., BONTCHEVA K., SCHARL A., FÖLS M., "Games with a Purpose or Mechanised Labour ? A Comparative Study", *i-Know '13 Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*, no. 19, p. 8, 2013.

[SAG 11] SAGOT B., FORT K., ADDA G. *et al.*, "Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé", *TALN 2011*, p. 12, Montpellier, 2011.

[SAR 14] SARROUY O., Faire foule. Organisation, communication et (dé)subjectivisation à l'ère hyperindustrielle, Thesis, University of Rennes 2, 2014.

[SAY 11] SAYLOR N., WOLFE J., "Experimenting with Strategies for Crowdsourcing Manuscript Transcription", *Research Library Issues: a quarterly report from ARL, CNI, and SPARC*, 2011.

[SCH 05] SCHULTZ P., "The Producer as Poweruser", in COX G., KRYSA J. (eds) *Engineering Culture: on the author as (digital) producer*, Autonomedia, New York, pp. 111–125, 2005.

[SCH 08] SCHULTZ P., "Market Ideology and the Myths of Web 2.0", *First Monday*, vol. 13, no. 3, 2008.

[SCH 10] SCHENK E., GUITTARD C., Le crowdsourcing: modalités et raisons d'un recours à la foule, INSA, Lyon, 2010.

[SCH 12] SCHENK E., GUITTARD C., "Une typologie des pratiques de Crowdsourcing : l'externalisation vers la foule, au-delà du processus d'innovation", *Management international*, vol. 16, pp. 89–100, 2012.

[SHA 10] SHARMA A., Crowdsourcing Critical Success Factor Model: Strategies to harness the collective intelligence of the crowd, Working paper, 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.

[SMI 13] SMITH D., MANESH M., ALSHAIKH A., "How Can Entrepreneurs Motivate Crowdsourcing Participants?", *Technology Innovation Management Review*, vol. 3, no. 2, pp. 23–30, 2013.

[SNO 08] SNOW R., O'CONNOR B., JURAFSKY D. *et al.*, "Cheap and Fast – But is it Good ? Evaluating Non-Expert Annotations for Natural Language Tasks", *Proceeding EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

[STA 13] STAMBAUGH E.K., "Reinventing Shared Print : A Dynamic Service Vision for Shared Print Monographs in a Digital World", *Against the Grain*, pp. 68–70, CDL Staff Publications, Oakland, 2013.

[STE 14] STEINBACH L., "Digital cultural heritage is getting crowded : crowdsourced, crowd-funded, and crowd-engaged", *Digital Heritage and Culture : Strategy and Implementation*, 2014.

[STI 14] STILLER J., From Curation to Collaboration A Framework for Interactions in Cultural Heritage Information Systems, Doctoral thesis, Humboldt University of Berlin, 2014.

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

[SUR 04] SUROWIECKI J., *The Wisdom of Crowds, Doubleday*, 2004

[SZO 12] SZONIECKY S., Evaluation et conception d'un langage symbolique pour l'intelligence collective : vers un langage allégorique pour le Web, Thesis, Paris 8 University, Vincennes-Saint Denis, 2012.

[TAF 11] TAFURI N., MAYS A., "Bullied by Budgets, Pushed by Patrons, Driven by Demand : Libraries and Tantalizing Technologies", *Proceedings of the Charleston Library Conference*, pp. 317–323, Charleston, United States, 2011.

[THO 12] THOGERSEN R., "Crowdsourcing for image metadata; a comparison between game-generated tags and professional descriptors", *Lecture Notes in Computer Science*, no. 8224, 2012.

[THU 13] THUAN N.H., ANTUNES P., JOHNSTONE D., "Factors Influencing the Decision to Crowdsource", *Lecture Notes in Computer Science*, no. 8224, pp. 110–125, 2013.

[TRA 06] TRANT J., WYMAN B., "Investigating social tagging and folksonomy in art museums with steve.museum", *Collaborative Web Tagging Workshop at WWW2006*, p. 6, Edinburgh, Scotland, 2006.

[TRA 08] TRAINOR C., Open Source, Crowd Source: harnessing the power of the people behind our libraries. Library Faculty and Staff Papers and Presentations, Paper 3, 2008.

[TWE 12] TWEDDLE J.C., ROBINSON L.D., POCOCK M.J.O. *et al.*, "Guide to citizen science: developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK", Centre for Ecology & Hydrology, available at: http://www.ceh.ac.uk, p. 29, 2012.

[VER 13] VERSHBOW B., "NYPL Labs : Hacking the Library", *Journal of Library Administration*, vol. 53, no. 1, pp. 10–26, 2013.

[VON 04] VON AHN L., DABBISH L., "Labeling Images with a Computer Game", *ACM Conf. on Human Factors in Computing Systems CHI*, Vienna, Austria, pp. 319–326, 2004.

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

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

[VON 06b] VON AHN L., LIU R., BLUM M., "Peekaboom: A Game for Locating Objects in Images", *Proceeding CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 55–64, Quebec, Canada, 2006.

[VON 06c] VON AHN L., KEDIA M., BLUM M., “Verbosity: A Game for Collecting Common-Sense Facts”, *Proceeding CHI ‘06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 75–78, Quebec, Canada, 2006.

[VON 08a] VON AHN L., DABBISH L., “Designing Games With A Purpose”, *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[VON 08b] VON AHN L., MAURER B., MCMILLEN C. *et al.*, “reCAPTCHA : Human-Based Character Recognition *via* Web Security Measures”, *Science*, no. 321, pp. 1465–1468, 2008.

[VON 11] VON HIPPEL E., OGAWA S., DE JONG J.P.J., “The age of the Consumer-Innovator”, *MIT Sloan Management Review*, vol. 53, no. 1, pp. 27–35, 2011.

[WAI 08] WAIBEL G., “You’re More Social Than You Think?”, *Presentation, GNCTPG Annual Meeting*, Alameda Public Library, United States, 2008.

[WEN 98] WENGER E., *Communities of Practice: Learning, Meaning, and Identity*, Cambridge University Press, Cambridge, 1998.

[WIL 11] WILSON-HIGGINS S., “Could print on demand actually be the “new interlibrary loan”?”, *Interlending & Document Supply*, vol. 39, no. 1, pp. 5–8, 2011.

[ZAR 14] ZARNDT F., “Crowdsourcing family history, and long tails for libraries”, *IFLA Newspaper Section*, p. 83, 2014.

[ZAS 14] ZASTROW J., “Crowdsourcing Cultural Heritage: “Citizen Archivists” for the Future”, *The Digital Archivist*, vol. 34, no. 8, p. 4, 2014.

Index

Digital Libraries and Crowdsourcing, First Edition. Mathieu Andro.

Other titles from

iSTE

in

Information Systems, Web and Pervasive Computing

2017

BOUHAÏ Nasreddine, SALEH Imad
Internet of Things: Evolutions and Innovations
(Digital Tools and Uses Set – Volume 4)

DUONG Véronique
Baidu SEO: Challenges and Intricacies of Marketing in China

LESAS Anne-Marie, MIRANDA Serge
The Art and Science of NFC Programming
(Intellectual Technologies Set – Volume 3)

LIEM André
Prospective Ergonomics
(Human-Machine Interaction Set – Volume 4)

MARSAULT Xavier
Eco-generative Design for Early Stages of Architecture
(Architecture and Computer Science Set – Volume 1)

REYES-GARCIA Everardo
The Image-Interface: Graphical Supports for Visual Information
(Digital Tools and Uses Set – Volume 3)

REYES-GARCIA Everardo, BOUHAÏ Nasreddine
Designing Interactive Hypermedia Systems
(Digital Tools and Uses Set – Volume 2)

SAÏD Karim, BAHRI KORBI Fadia
Asymmetric Alliances and Information Systems:Issues and Prospects
(Advances in Information Systems Set – Volume 7)

SZONIECKY Samuel, BOUHAÏ Nasreddine
Collective Intelligence and Digital Archives: Towards Knowledge Ecosystems
(Digital Tools and Uses Set – Volume 1)

2016

BEN CHOUIKHA Mona
Organizational Design for Knowledge Management

BERTOLO David
Interactions on Digital Tablets in the Context of 3D Geometry Learning
(Human-Machine Interaction Set – Volume 2)

BOUVARD Patricia, SUZANNE Hervé
Collective Intelligence Development in Business

EL FALLAH SEGHROUCHNI Amal, ISHIKAWA Fuyuki, HÉRAULT Laurent, TOKUDA Hideyuki
Enablers for Smart Cities

FABRE Renaud, in collaboration with MESSERSCHMIDT-MARIET Quentin, HOLVOET Margot
New Challenges for Knowledge

GAUDIELLO Ilaria, ZIBETTI Elisabetta
Learning Robotics, with Robotics, by Robotics
(Human-Machine Interaction Set – Volume 3)

HENROTIN Joseph
The Art of War in the Network Age
(Intellectual Technologies Set – Volume 1)

KITAJIMA Munéo
Memory and Action Selection in Human–Machine Interaction
(Human–Machine Interaction Set – Volume 1)

LAGRAÑA Fernando
E-mail and Behavioral Changes: Uses and Misuses of Electronic Communications

LEIGNEL Jean-Louis, UNGARO Thierry, STAAR Adrien
Digital Transformation
(Advances in Information Systems Set – Volume 6)

NOYER Jean-Max
Transformation of Collective Intelligences
(Intellectual Technologies Set – Volume 2)

VENTRE Daniel
Information Warfare – 2nd edition

VITALIS André
The Uncertain Digital Revolution

2015

ARDUIN Pierre-Emmanuel, GRUNDSTEIN Michel, ROSENTHAL-SABROUX Camille
Information and Knowledge System
(Advances in Information Systems Set – Volume 2)

BÉRANGER Jérôme
Medical Information Systems Ethics

BRONNER Gérald
Belief and Misbelief Asymmetry on the Internet

IAFRATE Fernando
From Big Data to Smart Data
(Advances in Information Systems Set – Volume 1)

KRICHEN Saoussen, BEN JOUIDA Sihem
Supply Chain Management and its Applications in Computer Science

NEGRE Elsa
Information and Recommender Systems
(Advances in Information Systems Set – Volume 4)

POMEROL Jean-Charles, EPELBOIN Yves, THOURY Claire
MOOCs

SALLES Maryse
Decision-Making and the Information System
(Advances in Information Systems Set – Volume 3)

SAMARA Tarek
ERP and Information Systems: Integration or Disintegration
(Advances in Information Systems Set – Volume 5)

2014

DINET Jérôme
Information Retrieval in Digital Environments

HÉNO Raphaële, CHANDELIER Laure
3D Modeling of Buildings: Outstanding Sites

KEMBELLEC Gérald, CHARTRON Ghislaine, SALEH Imad
Recommender Systems

MATHIAN Hélène, SANDERS Lena
Spatio-temporal Approaches: Geographic Objects and Change Process

PLANTIN Jean-Christophe
Participatory Mapping

VENTRE Daniel
Chinese Cybersecurity and Defense

2013

BERNIK Igor
Cybercrime and Cyberwarfare

CAPET Philippe, DELAVALLADE Thomas
Information Evaluation

LEBRATY Jean-Fabrice, LOBRE-LEBRATY Katia
Crowdsourcing: One Step Beyond

SALLABERRY Christian
Geographical Information Retrieval in Textual Corpora

2012

BUCHER Bénédicte, LE BER Florence
Innovative Software Development in GIS

GAUSSIER Eric, YVON François
Textual Information Access

STOCKINGER Peter
Audiovisual Archives: Digital Text and Discourse Analysis

VENTRE Daniel
Cyber Conflict

2011

BANOS Arnaud, THÉVENIN Thomas
Geographical Information and Urban Transport Systems

DAUPHINÉ André
Fractal Geography

LEMBERGER Pirmin, MOREL Mederic
Managing Complexity of Information Systems

STOCKINGER Peter
Introduction to Audiovisual Archives

STOCKINGER Peter
Digital Audiovisual Archives

VENTRE Daniel
Cyberwar and Information Warfare

2010

BONNET Pierre
Enterprise Data Governance

BRUNET Roger
Sustainable Geography

CARREGA Pierre
Geographical Information and Climatology

CAUVIN Colette, ESCOBAR Francisco, SERRADJ Aziz
Thematic Cartography – 3-volume series
Thematic Cartography and Transformations – Volume 1
Cartography and the Impact of the Quantitative Revolution – Volume 2
New Approaches in Thematic Cartography – Volume 3

LANGLOIS Patrice
Simulation of Complex Systems in GIS

MATHIS Philippe
Graphs and Networks – 2nd edition

THERIAULT Marius, DES ROSIERS François
Modeling Urban Dynamics

2009

BONNET Pierre, DETAVERNIER Jean-Michel, VAUQUIER Dominique
Sustainable IT Architecture: the Progressive Way of Overhauling Information Systems with SOA

PAPY Fabrice
Information Science

RIVARD François, ABOU HARB Georges, MERET Philippe
The Transverse Information System

ROCHE Stéphane, CARON Claude
Organizational Facets of GIS

2008

BRUGNOT Gérard
Spatial Management of Risks

FINKE Gerd
Operations Research and Networks

GUERMOND Yves
Modeling Process in Geography

KANEVSKI Michael
Advanced Mapping of Environmental Data

MANOUVRIER Bernard, LAURENT Ménard
Application Integration: EAI, B2B, BPM and SOA

PAPY Fabrice
Digital Libraries

2007

DOBESCH Hartwig, DUMOLARD Pierre, DYRAS Izabela
Spatial Interpolation for Climate Data

SANDERS Lena
Models in Spatial Analysis

2006

CLIQUET Gérard
Geomarketing

CORNIOU Jean-Pierre
Looking Back and Going Forward in IT

DEVILLERS Rodolphe, JEANSOULIN Robert
Fundamentals of Spatial Data Quality