Digital Invasions Within Cultural Heritage: Social Media and Crowdsourcing

Lorenzo Monti[1(✉)], Giovanni Delnevo[1], Silvia Mirri[2], Paola Salomoni[2], and Franco Callegati[1]

[1] Interdepartmental Centre for Industrial ICT Research - CIRI ICT, Università di Bologna, Bologna, Italy
lorenzo.monti20@unibo.it
[2] Department of Computer Science and Engineering, Università di Bologna, Bologna, Italy

Abstract. The wide diffusion of mobile devices and of digital technologies are dramatically changing the usage scenarios in different contexts. One of them is cultural heritage, where new media are offering huge possibilities for the enhancement and the enrichment of heritage experience, improving the users' involvement. In particular, tourists equipped with their mobile devices are invading cultural attractions, sharing pictures and comments (together with hashtags and geo-localized positions) on social networks. These represent a source of data, which can be integrated with the official ones provided by GLAM (Galleries, Libraries, Archives, and Museums) and cultural heritage institutions, enriching them. In this paper, we explore how social networks and crowdsourcing activities can be exploited as a source of information for cultural places and pieces of art.

Keywords: Web scraping · Data extraction · Open data
Cultural heritage · Crowdsourcing · Social network
Crowdsourced data · Social media data

1 Introduction

Society is moving toward the post-industrial age, where the consumer, as we know, no longer exists and s/he is replaced with a new kind of user that is not just a mere consumer, but also a producer of contents [1]. These new users are called *prosumer*, term branded by Toffler [2]. From this perspective, there will be new professionals and there will be an unavoidable metamorphosis from the actual ones, driven by the rise of resilient factors mainly from the self-adjustment to new job needs. The first half of twentieth century was certainly marked by a huge technological change. An example of this change is the advent of cinematography, that have certainly been an impact also in the way to interpret and experience cultural heritage. One of the most important dissertation about

B. Guidi et al. (Eds.): GOODTECHS 2017, LNICST 233, pp. 102–111, 2018.
https://doi.org/10.1007/978-3-319-76111-4_11

this theme was written by Benjamin, in his famous work 'The Work of Art in the Age of Mechanical Reproduction' in [3]. He claims that the introduction of a new technique to produce, reproduce and spread worldwide artworks, has radically changed the attitude towards art of both artists and public [4]. The author focuses on the particular kind of cultural heritage, that is artwork which, through mechanical reproduction techniques like films rather than printing, destroyed the concept of "aura" of an artwork. The aura is broadly understood as the feeling of religious nature resulted in the spectator in front of the original specimen of an artwork. As with the advent of mechanical reproduction, new digital media have radically altered the concept of cultural heritage: digitization technique, animation, 3D reconstruction or immersive learning are only some examples and have certainly changed and redefined the concept of transmission and use of knowledge [5]. The use of multi-touch monitors used like information tools, the realization of consoles with reconstructions of three dimensional digital model, visualizations of active/passive anaglyph, immersive augmented reality, synchrony or diachrony information analysis or applications for low-vision users based on touch tactile system are sample scenarios of media that promote a different interaction with knowledge so as to improve visitors' engagement [6]. In Web applications, the progressive technological development has led to the creation of different platforms with expanded visualization and navigation capabilities of three-dimensional data more and more pertinent to the perceptual image quality and able to involve the user both in space and time terms [7].

New media offers enormous possibilities for the enhancement and enrichment of heritage experience and interpretation. In the urban context, municipalities and public entities should provide adequate infrastructure, so as to guarantee distribution and sharing of multimedia resources, such as audio and video [8]. The question is how to make best use of new media to maintain the integrity of heritage artifacts and sites. How this has to be achieved can vary according to particular heritage contexts, artifacts and sites, and it can also differ according to various curatorial practices and different media [9].

In these area lies the concept of *crowdsourcing*, a neologism used for the first time by Howe in [10]. It is formed by two words, *crowd* and *outsourcing* that indicates the outsourcing of data supply. There is no unique definition of crowdsourcing, but thanks to the work of Estellés-Arolas and González-Ladrón-De-Guevara [11], that tries to collect all the definitions of it and tries to identify the common characteristics of them, we can define as "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." Indeed, the increasing diffusion of mobile devices have contributed to the use of crowdsourcing also in geo-spatial context [12], changing the way the informations are produced, used and stored [13].

Tourists equipped with their smart phones are invading cultural attractions, taking pictures and selfies, sharing posts, hashtags [14], and their geo-localized position, tagging friends and places. In this way, they are populating and enriching social networks with such digital media (sometimes within specific events planned or supported by cultural institutions or by social media communities, such as the Wiki Loves Monuments[1] and the Digital Invasions[2] initiatives). These new contents could be exploited by cultural institutions, tourism offices, public entities, and private foundations, with the aim of monitoring activities related to their goods and places on social media. In particular, the use and the collection of pictures coming from social networks, from non-common devices (i.e. vehicles on board cameras [15]), from IoT and cloud contexts can play a significant role.

With this goal in mind, in this paper we present an experiment we have conducted, with the aim of gathering, grouping, and evaluating media and content coming from crowdsourcing activities in social media, related to cultural heritage (i.e. historical palazzos, monuments, museums, pieces of art, etc.), by using scraping techniques.

The remainder of the paper is structured as follows. Section 2 provides a brief description of the background and presents some related work. Section 3 summarizes the approach we have exploited. Section 4 describes the architecture and the implementation of the system, while the results are illustrated and discussed in Sect. 5. Finally, Sect. 6 concludes the paper highlighting some final remarks and future work.

2 Background and Related Work

Since the late 90s' the potential of new media has improved many areas and among them there cultural heritage. This has occurred in various forms. An example is the reconstruction, based on 3D modeling, of archaeological sites or lost monuments and cultural heritage due to war conflicts (just like Palmyra site in Syria) or due to natural disasters (just like tsunamis or earthquake [16]). Thanks to crowdsourcing and to motion technique, used for the reconstruction of a 3D model, it has been possible to reconstruct a model of Plaka Bridge, a 19th-century stone one-arch bridge in Greece, that collapsed during the floods of 1 February 2015. In this project, acquired images are uploaded from different kind of users, with the result that they have different resolution, perspective, distance and brightness. Thus, it has to be considered that they are also snapped in different seasons and times.

Researchers also investigated how transcription is even more important, in historical documents, than the uniqueness of most of these documents and the preservation of their contents is essential for historical and cultural reasons. Transcribing handwritten documents through OCR technologies isn't always enough and is far from perfect. Crowdsourcing and human experts' revision emerged as a powerful tool in order to obtain a correct transcription. Through the rise of a

[1] https://www.wikilovesmonuments.org/.
[2] http://www.invasionidigitali.it/en/.

crowdsourcing platform [18], in which users contribute to a small given task, getting back a little or null payment, the transcription of cultural texts has spread. We can find examples in tranScriptorium [19], an European project that aims to develop a solution for annotating handwritten historical documents using modern, holistic Handwritten Text Recognition (HTR) technology, or alternative solutions, such as the use of speech dictation of handwritten text lines as transcription source in a crowdsourcing platform [20]. Another use of crowdsourcing in cultural heritage context regards tagging or captioning images. Despite advances in the field of content-based image retrieval, human intake recovers the higher semantic level within cultural heritage image databases, gradually shifting users from passive consumer to pro-active users (manipulating data, improving information retrieval, etc.) [21].

Crowdsourcing is a flexible model that can be applied to a wide range of activities among which adding content to maps like in [22], where it has been developed an application that suggests LPOI (Local Point Of Interest) around the user, through geolocation. Flickr was used to retrieve resources and Wikipedia was used in order to add information. Once the LPOI are chosen, the system computes and displays the suggested path. Every photo retrieved from Flickr is consider as "candidate" until it reached a majority vote, thanks to users' feedback. Then, such a picture becomes eligible of being (or not) displayed whenever a search result encompasses the related LPOI. The purpose of this project is to get users' feedback to improve results displayed by the system.

In the cultural heritage context, the use of digital technologies is increasingly frequent with the aim of enlarging the use of information based on multisensory and multimodal interaction mechanisms and involving user in the exploration of content in a proactive way. An example is 'Youtube Play', a project promoted by Solomon R. Guggenheim, Youtube and HP: some videos selected from a contest were projected from 22 to 24 October 2010 on the facade and in the inner round of the Guggenheim of New York [23]. A similar project was the CITYCLUSTER one [24] that is based on a virtual-reality networking matrix, where interactivity, graphic, and content style coexist in a common virtual territory. Such a project is connected through a high-speed network in which shared environments (both real and imagined) enable remote participants to collaborate, interact, and work together in a common virtual space over distance in real time.

In this context, another concept that can play a key role is gamification [25]. In particular, gamification of micro-tasks spur people coming back for more [26]. Obviously, the main goals of crowdsourcing concern the possibility to create innovative solutions or saving money, but organize required activities more like playing a game can help to remove concerns from people who make tasks for free or for small payment. In fact, gamification has a positive impact on crowdsourcing, both from qualitative and qualitative points of view [26,27].

3 Our Approach

In the context of cultural heritage within a urban area, there are different types of data. On the one hand there are official data coming from GLAM (Galleries,

Libraries, Archives, Museums), public administrations, catalogs, and foundations. On the other hand, there are private citizens who, thanks to web 2.0, social networks, and crowdsourcing, have become *prosumer*, as being provider and consumer at the same time, adding contents (photos, tag, feedback, comments) on the Web. Therefore, a large quantity of data that can be found on crowdsourcing platforms, which could support and enrich information already given by official entities. In the context of the SACHER project (Smart Architecture for Cultural Heritage in Emilia Romagna[3], which is co-funded by the Emilia Romagna Region through the POR FESR 2014–2020 fund - European Regional Development Fund), we aim to integrate such two categories of data so as to provide immediate information to users who otherwise would have to do a much longer research. In this paper, we present a first prototype devoted to verify the possibility to gather information from some different cultural heritage sources, with specific regards to social media.

We exploit an open data WebGIS ("Patrimonio culturale dell'Emilia Romagna", that is the Italian for Cultural Heritage of Emilia Romagna region), which lets users visualize a map with points of interest related to cultural heritage (i.e. museums, palazzos, churches, monuments, etc.) and their relative information made available by the Segretariato Regionale for Emilia Romagna region of MiBACT (Minister of cultural heritage and activities and tourism). These open data include geo-localized architectural and archaeological goods. This WebGIS is constantly updated and it is growing: at the time we are writing, it includes 9,092 points of interest.

As crowdsourced data, we take into account data coming from social networks, mainly following two different approaches: *API*, provided directly by the social network considered and *Ad-hoc script*, when official APIs are not provided or if they too limited to be efficiently used. In particular, we aim to select and collect images shared on social media, recording title, description, and hashtags for each one. Our idea is to regularly acquire new images related to some specific case studies, uploaded by users on the different social networks we are monitoring, in order to test the feasibility of scraping those kinds of content from such platforms.

4 Implementation

The architecture of this project is designed bearing in mind the possibility of using both public and open data, like data from public administrations and GLAM and crowd-data from social media, gathering and integrating together to enrich and improve the information available to any users about specific cultural places and pieces of art. The experiment has been designed by considering the following different components, depicted in Fig. 1:

– *Open data:* these data are provided from API offered by different sources like public administrations, GLAM, etc. We stored them into database located in our server.

[3] http://www.eng.sacherproject.com/.

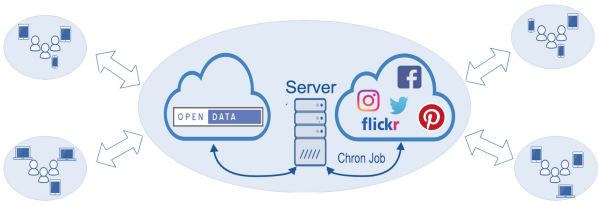

Fig. 1. Software architecture

- *Crowd data:* these data are provided from scraping Python scripts run through a chron-job, keeping continuously updated data from social media, such as Facebook and Twitter.
- *Server:* the server receives data and stores them in a dedicated database. The server is also used in order to reply to users' requests.

In particular, this work focuses on crowd data taken down from five of the most influent and commonly used social networks: Facebook, Twitter, Instagram, Pinterest, and Flickr. In order to develop scraping scripts, Python 2.x and 3.x have been used. Some official APIs (provided by social networks) have been exploited, as well as some specific Python modules, including:

- *Requests v2.18.2*[4], it is a simple module that allows to send HTTP/1.1 requests. It is possible to add parameters, form data, headers, multi-part files with simple Python dictionaries and to access the response data in the same way.
- *Urllib and Urllib2*[5], they are similar to the previous one, but they are a part of the Python Standard Library. Urllib and Urllib2 are modules that allow HTTP/1.1 requests.
- *BeautifulSoup v4.4*[6], it is a Python module for pulling data out of HTML and XML files. BeautifulSoup parses data acquired by Requests or Urllib/Urllib2 modules, which allows searching, navigating, and modifying the parsed tree.
- *Selenium Webdriver v3.4.0*[7], it is a browser automation tool, which automates the control of a browser, so that repetitive tasks can be automated. It has been used together with BeautifulSoup to parse and therefore automate scraping phases.
- *Tkinter*[8], it allows to create GUI in order to manage every scraper scripts.

Each Python script manages to find images and their descriptions as the result of a specific query. In particular, the scripts we have implemented are devoted to collect information about two of the most famous monuments in the city of

[4] https://github.com/requests/requests.
[5] https://docs.python.org/3.6/library/urllib.html#module-urllib.
[6] https://github.com/newvem/beautifulsoup.
[7] https://github.com/SeleniumHQ/selenium.
[8] https://docs.python.org/3.6/library/tkinter.html#module-Tkinter.

Bologna: “Fontana del nettuno” and “Palazzo del podestà”. Furthermore, data were collected at regular intervals in order to investigate how social networks could be exploited as a source of crowdsourced information about cultural heritage.

5 Results and Discussion

In this work, we have taken into account and monitored crowdsourced information about “Fontana del nettuno” and “Palazzo del podestà”, within the following social networks:

- *Flickr:* it was used Flickrapi[9], a native API from Flickr; in this way it was possible to scrape and download a large amount of images, as shown in Figs. 2 and 3;
- *Facebook:* thanks to Facepy[10], a Python library that makes easy to interact with Facebook’s Graph API. We managed to scrape images though in limited manner.
- *Twitter:* TwitterSearch API[11] are used in order to download data from the result of a specific query.
- *Instagram:* it was created an ad hoc Python script that uses Selenium Webdriver library. Selenium takes control of the browser in order to automatize descriptions and images download.
- *Pinterest:* it was implemented an hoc Python script that uses BeautifulSoup and Requests libraries. Pages are first requested, then their DOMs are parsed in order to find all the needed elements.

We scraped images for twelve consecutive days, starting from the 12th of June, until the 24th of June. We searched all the posts/tweets/pins/photos with the hashtags #fontanadelnettuno and #palazzodelpodestà. Results related to the hashtag *fontanadelnettuno* scraped from each monitored social network are reported in Fig. 2, while data shown in Fig. 3 are related to the results obtained for the hashtag *palazzodelpodestà*. In both cases, the social networks with most images are Instagram and Flickr along the monitored period. This is not surprising, since they are platforms mainly devoted to share pictures. The other analyzed social networks reported results pretty far from these two ones.

Some considerations arose from this first experiment and its results:

- We searched items with a particular hashtag and this could be limiting. A first improvement could be searching for the name of the monument (or of the piece of art) also in the title and in the description. Moreover, the geographic coordinates could be considered.

[9] https://www.flickr.com/services/api/.
[10] https://github.com/jgorset/facepy.
[11] https://github.com/ckoepp/TwitterSearch.

– Palazzo del podestà has a final accent on the "a" letter. Users could use hashtags without it. So, for each monument could be searched hashtags with its name put in different forms, for example without the accent or without the articulated prepositions "del". Hence, limits related to languages can occur.
– The last consideration regards the uniqueness of the names of the monuments. In fact, there is also a Fontana del Nettuno in Florence and there are other Palazzo del Podestà in many Italian cities. In these cases a filter on the geographic coordinates could be very useful.

	Flickr	Facebook	Twitter	Instagram	Pinterest
12-Jun	4263	12	244	6861	1250
13-Jun	4263	12	244	6865	1276
14-Jun	4263	12	244	6866	1278
15-Jun	4263	12	244	6868	1279
16-Jun	4263	13	244	6868	1279
17-Jun	4311	13	244	6871	1279
18-Jun	4311	13	244	6876	1290
19-Jun	4315	13	244	6881	1290
20-Jun	4318	13	244	6883	1294
21-Jun	4318	13	244	6886	1294
22-Jun	4332	13	244	6887	1311
23-Jun	4340	13	244	6900	1313
24-Jun	4357	13	244	6902	1315

Fig. 2. Fontana del Nettuno's scraping results: total number of pictures retrieved on a given day in a given data source (Flickr, Facebook, Twitter, Instagram, Pinterest)

	Flickr	Facebook	Twitter	Instagram	Pinterest
12-Jun	373	122	69	877	112
13-Jun	373	122	69	877	112
14-Jun	374	122	69	877	115
15-Jun	374	122	69	877	121
16-Jun	374	122	69	877	121
17-Jun	378	122	70	877	128
18-Jun	378	122	70	877	128
19-Jun	378	122	71	879	128
20-Jun	378	122	71	881	137
21-Jun	378	122	71	882	139
22-Jun	378	122	71	882	139
23-Jun	378	122	71	882	144
24-Jun	380	122	71	883	146

Fig. 3. Palazzo del podestà's scraping results: total number of pictures retrieved on a given day in a given data source (Flickr, Facebook, Twitter, Instagram, Pinterest)

6 Conclusions and Future Works

This first experiment confirms that social networks could be a useful source of information and media in the cultural heritage field, hence the could be used as a crowdsourcing platform and as a data source, to be integrated with other

official ones. This work highlights the non-feasibility of the ad hoc scripts in a long term perspective. In fact, a month later the tests, some of the ad hoc scripts fails, because the DOM was changed. Despite this, it is the only possible approach where the APIs are not provided or are too limited. In the first case, however, it's not long-term sustainable. In the second one, instead, an hybrid approach could be use. A first scraper could be made through the use of an ad-hoc script. Then APIs, albeit limited, could be used cyclically to get the new updates since the daily amount of data can be managed.

Interesting future works may include the use of techniques of image recognition with a double purpose. The first one is to understand if the monument is visible or not in the shared pictures (tagged with the related hashtags). The second one is to detect faces in order to crop or hide them for privacy reasons. In the perspective of crowdsourcing, a sentiment analysis could be done on descriptions of the photos, so as to get a raw evaluation about the visitors' perception and feelings about each monument.

References

1. Roccetti, M., Ferretti, S., Palazzi, C.E., Salomoni, P., Furini, M.: Riding the web evolution: from egoism to altruism. In: Proceedings of IEEE Consumer Communications and Networking Conference (CCNC 2008), pp. 1123–1127. IEEE (2008)
2. Toffler, A.: The Third Wave. William Morrow and Company. Inc., New York (1980)
3. Benjamin, W.: The Work of Art in the Age of Mechanical Reproduction of Art. Illuminations, Essays and Reflections (1936)
4. Dusi, N., Ferretti, I., Furini, M.: A transmedia storytelling system to transform recorded film memories into visual history. Entertain. Comput. **21**, 65–75 (2017)
5. Rosner, D., Roccetti, M., Marfia, G.: The digitization of cultural practices. Commun. ACM **57**(6), 82–87 (2014)
6. Mirri, S., Prandi, C., Roccetti, M., Salomoni, P.: Handmade narrations: handling digital narrations on food and gastronomic culture. J. Comput. Cult. Herit. (JOCCH) **10**(4), 20 (2017). Article No. 20
7. Guarnieri, A., Pirotti, F., Vettore, A.: Cultural heritage interactive 3D models on the web: an approach using open source and free software. J. Cult. Herit. **11**(3), 350–353 (2010)
8. Ghini, V., Salomoni, P., Pau, G.: Always-best-served music distribution for nomadic users over heterogeneous networks. IEEE Commun. Mag. **43**(5), 69–74 (2005)
9. Ott, M., Pozzi, F.: Towards a new era for cultural heritage education: discussing the role of ICT. Comput. Hum. Behav. **27**(4), 1365–1371 (2011)
10. Howe, J.: The Rise of Crowdsourcing. su Wired, giugno (2006)
11. Estellés-Arolas, E., González-Ladrón-De-Guevara, F.: Towards an integrated crowdsourcing definition. J. Inf. sci. **38**(2), 189–200 (2012)
12. Heipke, C.: Crowdsourcing geospatial data. ISPRS J. Photogramm. Remote Sens. **65**(6), 550–557 (2010)
13. Mirri, S., Prandi, C., Salomoni, P., Callegati, F., Melis, A., Prandini, M.: A service-oriented approach to crowdsensing for accessible smart mobility scenarios. Mob. Inf. Syst. **2016** (2016)

14. Furini, M., Mandreoli, F., Martoglia, R., Montangero, M.: The use of hashtags in the promotion of art exhibitions. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 187–198. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_15
15. Gerla, M., Weng, J.T., Pau, G.: Pics-on-wheels: photo surveillance in the vehicular cloud. In: 2013 International Conference on Computing, Networking and Communications (ICNC), pp. 1123–1127. IEEE (2013)
16. Stathopoulou, E.K., Georgopoulos, A., Panagiotopoulos, G., Kaliampakos, D.: Crowdsourcing lost cultural heritage. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **2**(5), 295 (2015)
17. Kaufmann, N., Schulze, T., Veit, D.: More than fun and money. Worker motivation in crowdsourcing-a study on Mechanical Turk. In: AMCIS, vol. 11, no. 2011, pp. 1–11 (2011)
18. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. Commun. ACM **54**(4), 86–96 (2011)
19. Sánchez, J.A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R.M., Vidal, E., de Does, J.: tranScriptorium: a European project on handwritten text recognition. In: Proceedings of 2013 ACM Symposium on Document Engineering, pp. 227–228. ACM, September 2013
20. Granell, E., Martínez-Hinarejos, C.D.: A multimodal crowdsourcing framework for transcribing historical handwritten documents. In: Proceedings of 2016 ACM Symposium on Document Engineering, pp. 157–163. ACM, September 2016
21. Van Hooland, S.: From spectator to annotator: possibilities offered by user-generated metadata for digital cultural heritage collections. In: Proceedings of CILIP Conference on Immaculate Catalogues: Taxonomy, Metadata and Resource Discovery in the 21st Century, September 2006
22. Bujari, A., Ciman, M., Gaggi, O., Palazzi, C.E.: Using gamification to discover cultural heritage locations from geo-tagged photos. Pers. Ubiquit. Comput. **21**(2), 235–252 (2017)
23. Gladysheva, D., Verboom, J., Arora, P.: The art tube: strategies, perceptions and outcomes of museums' online video portals. Digit. Cult. Educ. **6**(4), 393–408 (2014)
24. Fischnaller, F., Hill, A.: CITYCLUSTER "from the renaissance to the megabyte networking age": a virtual reality and high-speed networking project. Presence **14**(1), 1–19 (2005)
25. Prandi, C., Nisi, V., Salomoni, P., Nunes, N.J.: From gamification to pervasive game in mapping urban accessibility. In: Proceedings of 11th Biannual Conference on Italian SIGCHI Chapter, pp. 126–129. ACM (2015)
26. Morschheuser, B., Hamari, J., Koivisto, J.: Gamification in crowdsourcing: a review. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 4375–4384. IEEE, January 2016
27. Prandi, C., Roccetti, M., Salomoni, P., Nisi, V., Nunes, N.J.: Fighting exclusion: a multimedia mobile app with zombies and maps as a medium for civic engagement and design. Multimed. Tools Appl. **76**(4), 4951–4979 (2017)
28. Yuen, M.C., King, I., Leung, K.S.: A survey of crowdsourcing systems. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE 3rd International Conference on Social Computing (SocialCom), pp. 766–773. IEEE, October 2011
29. Brabham, D.C.: Crowdsourcing as a model for problem solving: an introduction and cases. Convergence **14**(1), 75–90 (2008)