IFLA WLIC 2019 Athens

Submitted on: 12.08.2019

ComunidadBNE: crowdsourcing at the National Library of Spain

Elena Sánchez Nogales
Dissemination of Digital Content, Web and Social Media
Division of Digital Processes and Services
National Library of Spain
elena.sanchez@bne.es

Abstract:

The National Library of Spain (Biblioteca Nacional de España, BNE) recently launched ComunidadBNE *(BNECommunity), a crowdsourcing project for the enrichment of its digital collections and data resources. It was developed within the Library's global strategy and framework (*BNElab*) aimed at promoting access, public participation and engagement, use and reuse of the Library's data and digital resources.*

ComunidadBNE is conceived as a digital environment of collaboration between general public, specialists and librarians, providing tools for cooperative transcription, georeferencing, identification or tagging of images from the Library's digital collections, and also for the enrichment of the bibliographic and authorities catalogues. The platform was built on and developed as open source and also integrates external data sources.

This paper aims at sharing the technical basis, contribution and validation processes, initial results, strategy and future developments expected for the project. And how crowdsourcing has become an important tool for librarians and users to jointly enrich and contribute to re-create a common heritage, by building together this new Community.

Keywords: Crowdsourcing; National Libraries; Digital communities; Cooperative cataloguing; Data enrichment.

Introduction

> *"No one knows everything, everyone knows something [and] all knowledge resides in humanity; digitalisation and communication technologies must become central in this coordination of far flung genius".*
>
> (Brabham, 2008)

Crowdsourcing is not a new term anymore. Not certainly for libraries.

It was coined by Jeff Howe in 2006 in his article *The Rise of Crowdsourcing* published in the *Wired* magazine, and described as "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call", thus here identified with a form of outsourcing and connected to the business sector.

Since then, the concept of crowdsourcing has evolved, expanded and given hundreds of definitions while being related to different ideas, each one offering distinctive theoretical or methodological approaches: citizen science, wisdom of the crowds, peer production, open sourcing, collective intelligence, user-generated content, community-based participatory research.

The library field has been naturally associated (at least over recent years) with the idea of using social engagement techniques to create community and awareness, but crowdsourcing introduced a new dimension: creating a formal and structured process for information transactions and intellectual input, defining a common goal for data enrichment, sharing librarian description techniques, letting the community build and improve library catalogues, creating new ways of social interaction to make our information more accessible and rich.

It is clear that all of these aspects pose interesting challenges to our role as 'information authorities', but benefits have been made evident while more and more initiatives arise in our field: from the obvious need of external knowledge to improve our data, to the subtler idea of strengthening bonds with the community or creating a feeling of social ownership with regards to the library collections, by encouraging public collaboration.

Crowdsourcing has proved to be a straightforward and powerful resource for such big goals, and in the library domain we have already seen a good number of interesting examples of large scale initiatives involving collaborative microtasks such as tagging, transcription, text correction, error marking or data addition. The National Library of Australia showed up as a firm early adopter, with programs like the Australian Newspaper Digitisation Program[1] for text correction (2008-2009) or Picture Australia[2] for addition of images (2006-2009). More recent examples of successfully launched crowdsourcing projects are for instance those by the New York Public Library[3], the British Museum[4], the British Library[5] or the Library of Congress[6].

At the National Library of Spain: background and strategic framework

The National Library of Spain (Biblioteca Nacional de España, BNE) has undertaken a steep digital transformation over the last ten years. With a rapidly increasing collection of digitized items, new digital resources including web archive and digitally born content, and a solid

[1] https://www.nla.gov.au/content/newspaper-digitisation-program
[2] https://trove.nla.gov.au/general/australian-pictures-in-trove
[3] For example, *Emigrant City*, for transcription of twentieth-century real state records: http://emigrantcity.nypl.org/#/
[4] *Micropasts*: https://crowdsourced.micropasts.org/
[5] *Libcrowds*: https://www.libcrowds.com/
[6] *By the people*: https://crowd.loc.gov/?loclr=blogsig

work on linked data and semantic enrichment of the catalogue with datos.bne.es[7], the process has introduced significant internal considerations on how the Library complies with its mission and public service, how to relate with and engage society, and how new digital ecosystems may and should be used.

Within very few years, the Library has created through digital and web environments new forms of institutional cooperation; learned new narratives to tell and share heritage; experimented with more tools and ways to support learning, research and innovation. And, very importantly, the institution soon understood the need of new approaches for collaborative creation, where citizens might and should be engaged to bring new values to cultural heritage and thus help make it still socially relevant and present in the digital era.

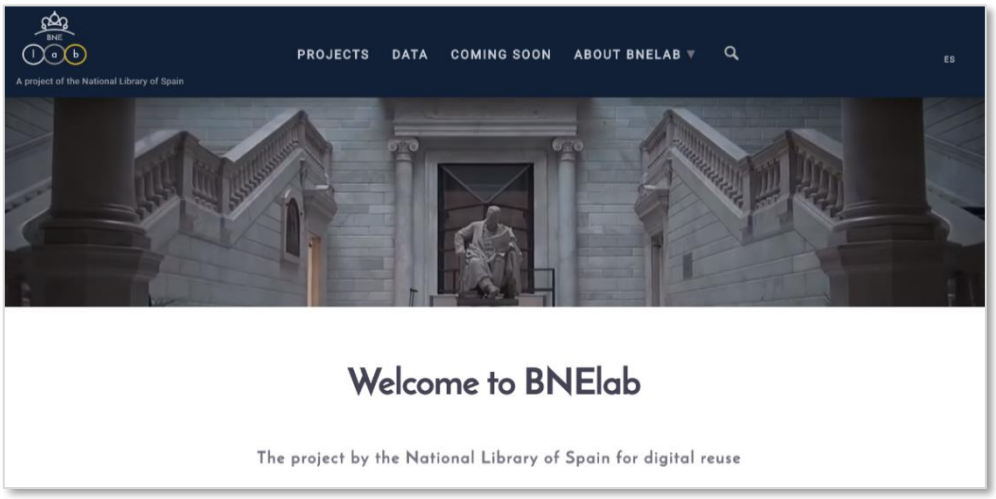
The path was early set as an institutional priority, and as such it was articulated as a strategic line (also in accordance with the legislation on the re-use of public sector information) in the 2015-2020 Strategy Plan[8].

In 2016, a comprehensive program was defined and launched with the support of Red.es, a public corporate entity under the Ministry of Energy, Tourism and the Digital Agenda[9].

The program covered different lines of action to promote the use and reuse of the BNE's data and digital resources:

- Data: semantic enrichment, creating new open and reusable datasets.
- New 'products' for inspiration and search for new uses of the digital collections (in gastronomy, design, fashion or tourism).
- Service platforms and digital environments, with tools and specific resources, for researchers, teachers or scholars.

BNElab[10] was set as the framework for all these initiatives.



[7] Datos.bne.es is the linked data portal of the BNE: http://datos.bne.es/inicio.html. Information about the project, model and technology in http://www.bne.es/en/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/index.html

[8] http://www.bne.es/webdocs/LaBNE/PlanEstrategico/Plan_estrategico_2015-2020.pdf [Spanish]

[9] Red.es is nowadays attached to the Ministry of Economy and Business.

[10] https://bnelab.bne.es/en/

Why a crowdsourcing project?

Crowdsourcing had always been a firm candidate to be included in this program. Similar projects had already been successfully launched by, for instance, the British Library, but why should the National Library of Spain work on a crowdsourcing initiative?

Main aspects considered were

- Making more and better data accessible and searchable in our catalogues (transcriptions, OCR correction or information not included in our records according to our description standards).
- Completing tasks that the library is not able to deal with, given the organization's resources and priorities (for instance again transcriptions or OCR correction).
- Enriching catalogues with expertise and knowledge from external communities.
- Showcasing new collections and materials.
- Giving visibility to our mission as a library and as librarians, by showing how our catalogues are built and how our library practice serves to describe and access collections.
- Involving and engaging society in building and enriching these catalogues for better access of present and future generations, thus creating a sense of common ownership on our heritage.
- Creating interest from new communities: groups with specific expertise or general public just willing to contribute to a non-profit cultural project.
- Building trust and bonds with the community: the Library asks for *help* and puts *trust* on contributors.

Major challenges identified were related to the standardization and validation of input, integration with the catalogue, and the need of having ways to boost and maintain participation in the platform.

Building ComunidadBNE

Being open source and access a primary principle within the BNElab program, finding a suitable, configurable and robust open source framework for the project was a key requirement. Well-founded solutions have been made available over the last years, being Scribe (used by the NYPL Labs) or Pybossa (the base of Libcrowds by the British Library, or Micropasts by the British Museum) two excellent examples.

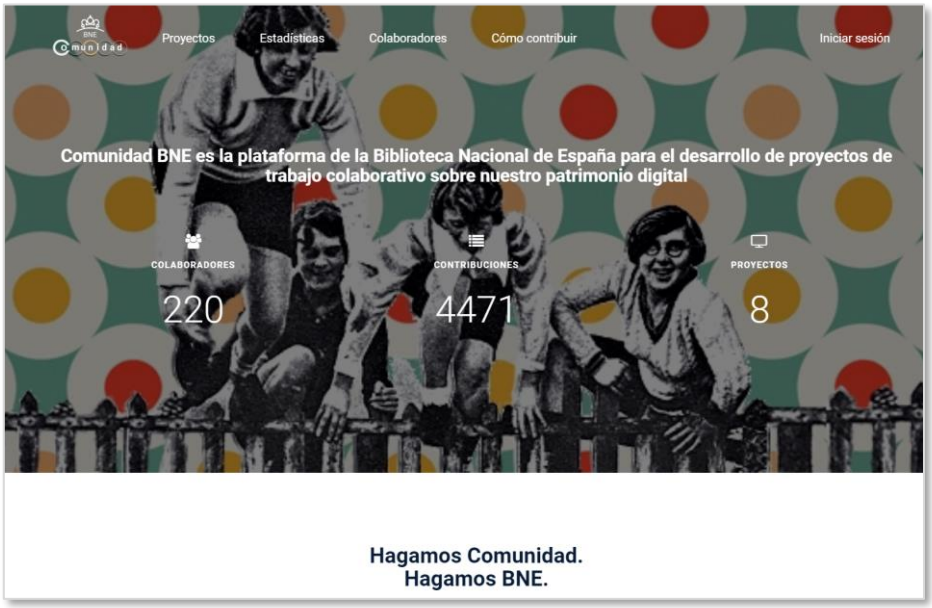
Pybossa[11] was developed by a Spanish company, Scifabric[12], and has proved to be a versatile and adaptable framework for many different types of crowdsourcing environments based on microtasking: research and heritage institutions, hospitals, universities or market research companies have used this technology for the development of platforms for data analysis and collaborative enrichment.

[11] https://pybossa.com/
[12] https://scifabric.com/

The technology code is available as open source in GitHub[13], with a good level of active development. It offers the basic framework as well as the template projects with the main components, from which additional features can be configured and customized. Pybossa allows integration with external data services; includes a statistics module and a robust validation system based on coincident contributions, easily configured for each task and project.

For ComunidadBNE, special features have been designed and developed for presentation of images, combination of different tasks and particularly for an efficient integration of the BNE's linked database and other data sources. The web interface was of course designed as well and adapted to the project's needs. All developments have been also made open and reusable in GitHub, under GNU General Public License[14].

The result in http://comunidad.bne.es/ shows an initial view of the number of projects and current contributions, and a simple menu for: current projects, statistics (basic data about projects, completed tasks and contributions), contributors (list and ranking) and information about the project. Users may work as registered users or anonymously.



Projects proposed for the launching in February 2019, and still active at the 'Proyectos' tag, try to show a variety of documents and types of enrichment, while illustrating the crowdsourcing features and possibilities developed and now available at the platform.

ComunidadBNE currently offers open contribution in these projects:

- *Unidentified*: identification of people and their personal stories, which lie unidentified in the vast collection of photographs from the Spanish Civil War.
- *Limelights:* transcription and identification of theatre persons and roles, from a collection of nineteenth-century posters.

[13] https://github.com/Scifabric/pybossa
[14] https://github.com/BNELab

- *Jean Laurent was here –and so he saw us*: location and georeferencing of photographs by Jean Laurent, an essential figure in the history of nineteenth-century photography in Spain.
- *To my distinguished friend...*: transcription of dedications on postcards and photographs.
- *Who is who*: identification of the portraits from important political figures contained in a work on the Constituent Assembly of 1869.
- *What does it sounds like?*: enrichment of authority records for musical groups (music genre and more).
- *A dictionary, a Swedish diplomat and nineteenth-century Spain*: marking, transcription and georeferencing of a peculiar manuscript where a diplomat described Spanish villages and their origins.
- *Translation by… [feminine, singular]*: from our catalogue records for the period 1900-1936, mentions to women translators have been extracted and proposed for enrichment or creation of authority records (with biographic and professional information).

The level of 'difficulty' is also intentionally varied, with projects requiring only transcription of a short text, while others implying several combined tasks and/or searching for information in external sources.

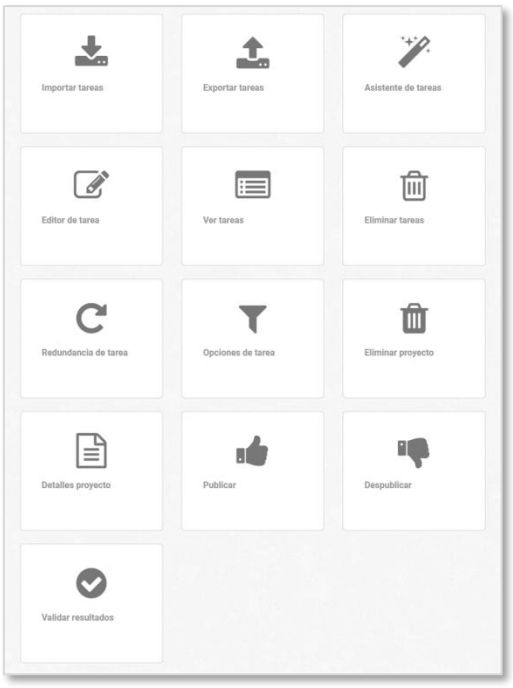
Creating and configuring a project in ComunidadBNE

ComunidadBNE offers a service platform for the creation of unlimited crowdsourcing projects through a rather straightforward web interface.

In fact, it could be said that the most important and difficult part of building a project comes *before* creating it in the platform: deciding on resources to be crowdsourced; analysing catalogue records and how they may be enriched; agreeing on the opportunity and convenience of potentially adding which data to the catalogue; determining vocabularies and authorities *against which* contributions will be made; studying how the resulting data would be incorporated into the records… These aspects can only be studied in close cooperation between cataloguing and technical/digital departments.

And in this regard, since the idea of a crowdsourcing project was put forward at the BNE, the initiative has proved to be an excellent exercise of internal collaboration and sharing of practices and views. As another evidence that new digital environments at the Library and 'traditional' cataloguing practice can and should always meet to enrich each other.

Once a project is designed, building it into ComunidadBNE is very simple. The interface smoothly guides the administrator through the steps of creating, describing and configuring.



Categories, interests or scope can be defined at a first step, as well as the possibility of allowing for anonymous contributions (by default, the platform permits non-registered users to participate) or make it a 'closed' project, meaning access is granted only with a password set for the project.

This option is particularly interesting for initiatives oriented to restricted communities or specifically designated users, either requiring special skills (in Paleography, for instance) or aimed to serve as a tool for research or academic groups.

Once the project template is selected from the ones available (for georeferencing, transcription, etc.), we may proceed to import data for the resources to be crowdsourced. This is done through a simple CSV file where the strictly needed information is just an ID number for the resource to be imported. Here, different types of resources have been defined, which will also mean different ways of presenting the resource to contributors. The ID may correspond to:

- a digital object, that will be imported from the BNE's digital repositories (Biblioteca Digital Hispánica[15] or Hemeroteca Digital[16]) and presented as an image to the users, or
- a catalogue resource (bibliographic or authority record) whose metadata will be called through an API from our linked data portal datos.bne.es.

Next step is configuring the tasks that will be suggested to users. This is done in a predefined JSON file where we will be adding all the texts, questions, answer options or vocabularies that will be presented for this project, as well as the number of coincident contributions that we will mark as required for each task before considered validated.

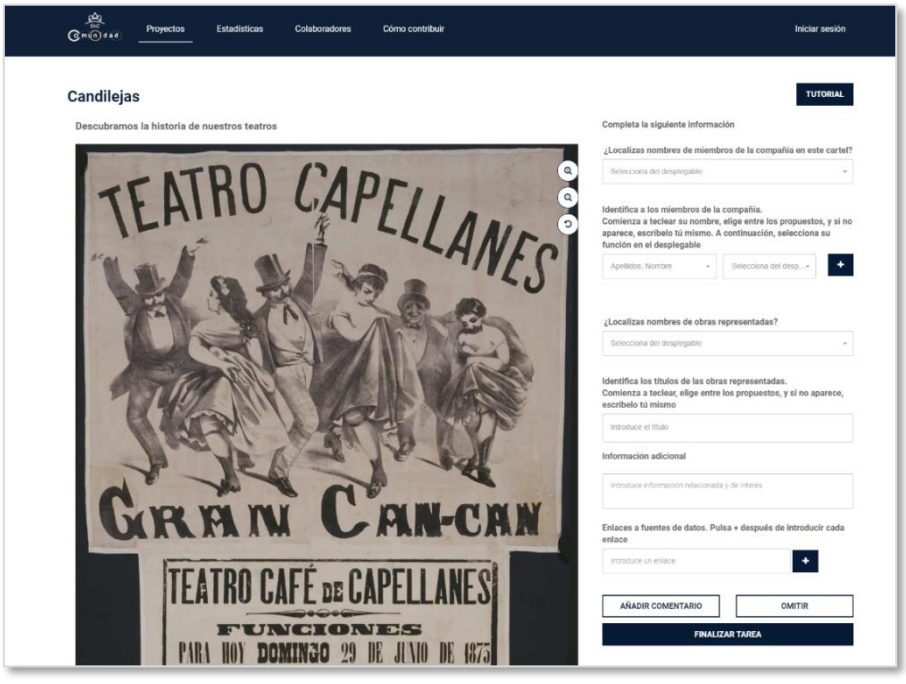
[15] http://bdh.bne.es
[16] http://hemerotecadigital.bne.es/index.vm

Not really being the most user-friendly part of the process, it is still quite straightforward to handle and certainly does not require for programming skills whatsoever.

Once completed, it is time to create a tutorial and publish the project.

Contributing to a project

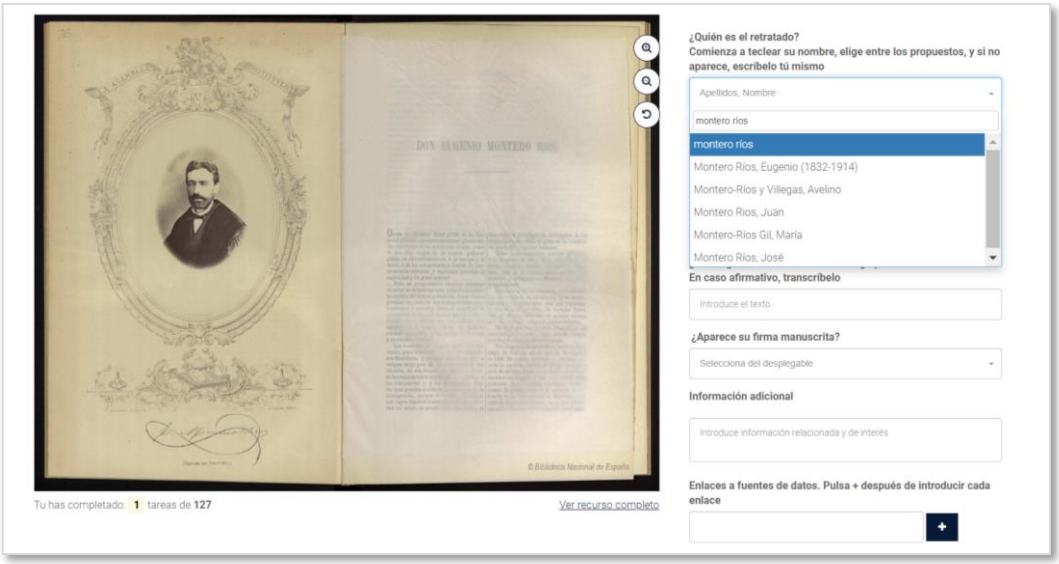
Tasks presentation and contributions panel is similar in all projects, with a left area where the image or catalogue record is displayed (images may be zoomed in or moved, and it is also possible to navigate from here to the original resource at the catalogue or digital library). On the right-hand side, tasks are presented to users.



Tasks may imply:

- Answering questions from image observation or knowledge/additional research: ‘Can you identify this person in the picture?’, ‘Does the image show an autograph?’, ‘Can you give this person’s place of birth/date of birth/occupation…?’
  Questions may be yes/no, closed (chosen from a limited vocabulary) or open –though this will obviously make validation more difficult–.

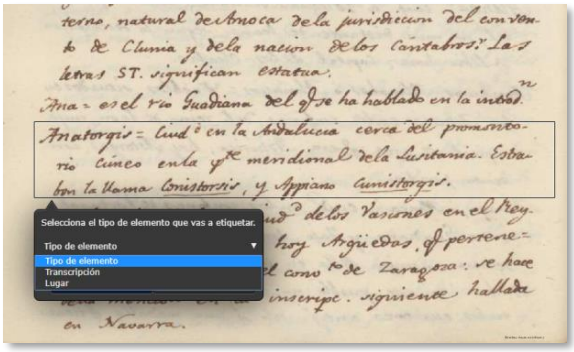
  Here, an important feature in ComunidadBNE is the integration of internal or external databases and vocabularies, importing or dynamically calling data through an API service. Thus, when we require the identification of a person, entity, work, subject… which may be available in our linked catalogue datos.bne.es, the contributor filling in the answer box will be dynamically suggested the options found in our catalogue records (via API), through auto-complete:



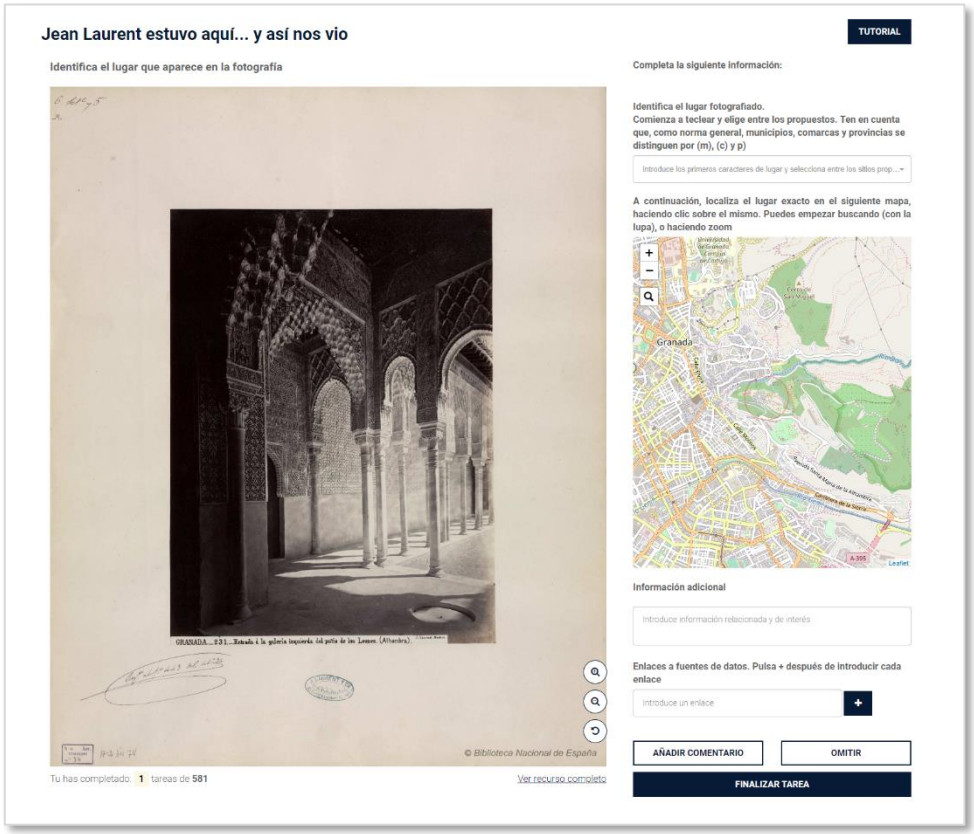
  Not only does this feature make contribution easier for the user, but also enables a more consistent validation of data for further integration into the catalogue records.

- Simple text transcription, to be included in an open question answer box.

- Marking a specific area in an image, and tagging the selected element as a place/person/work, or transcribing the text if required. For these cases, a special viewer has been developed to enable directly working on the image and the selected area.



- Identifying and georeferencing a place from an image or text. OpenStreetMap[17] service has been integrated here for searching and geolocating.



These are the basic types of tasks currently available, but all of them may be combined and personalized. A second development phase is currently on-going, and features are being improved (specially with regards to usability) or enhanced to allow more resource types and crowdsourcing options: audio-to-text transcription, TXT or XML files as project base for example for OCR (or OMR) correction, etc.

Validation and output

A project is considered automatically closed by ComunidadBNE when all suggested tasks are completed, since enough coincident contributions have been made and therefore auto-validated.

As mentioned earlier, the number of coincident contributions needed to consider an answer auto-validated is configured for each task, and this number will obviously depend on the nature of the task itself. For example, in a yes/no question we could get away with a low number of coincidences (in the range of 3-5), while the selection of a person from our authority records might require broader concurrence.

[17] https://www.openstreetmap.org

When are different answers deemed 'coincident' and therefore considered for auto-validation? For closed questions (yes/no or a selection from a drop-down list of options), it goes without saying that coincidence has to be total.

In the case of open questions such as text transcriptions, the auto-validation process combines two string-matching algorithms, Ratcliff/Obershelp and Levenshtein Distance, giving a match percentage. When the answers, compared and evaluated with this process, result in a 90% match or above (this can be configured), they are marked as candidates for auto-validation.

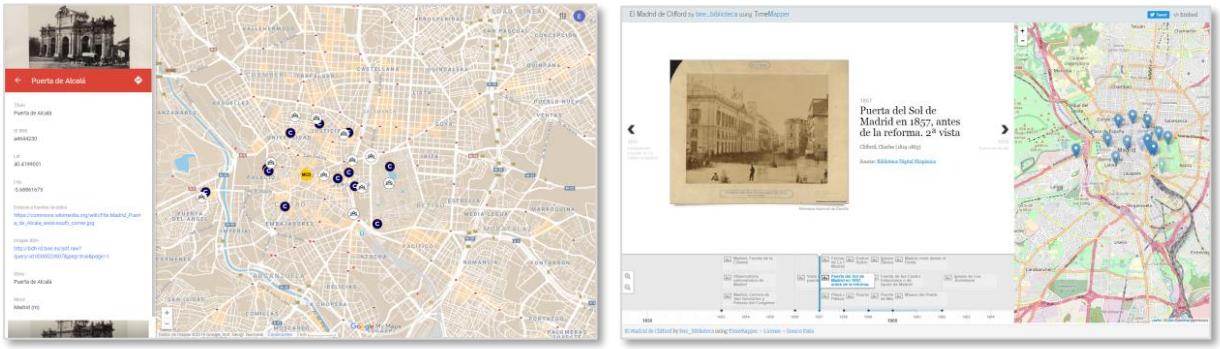
For georeferencing projects we designed a special validation process, considering the unlikeliness of having coincident exact same coordinates selected on the map by different contributors. A 'tolerance radius' is in these cases configured for each project, by setting a maximum distance (for example, 100 metres) to consider contributions coincident and therefore auto-validated.

The result of a completed project is a CSV file showing all the auto-validated contributions, which will be manually checked and validated by administrators. After that, data is ready for integration into the Library catalogue records, or for any further analysis, treatment or visualization.

This is an example of the resulting file from a georeferencing project (for a collection of nineteenth-century photographs of Madrid):

PID	ID BNE	Lat	Lng	Dist. entre p	Info adiciona	Enlaces a fuentes de datos	Imagen	Gbne	Mecd
bdh0000028	a4570852	40,4562215	-3,59812975	3.63 m			http://bdh-rd.bne.es/pdf.raw?qu	El Capricho (Madrid)	Madrid (m)
bdh0000023	a4643361	40,4178825	-3,71282905	52.01 m		https://es.m.wikipedia.or	http://bdh-rd.bne.es/pdf.raw?qu	Plaza de Oriente (Madrid)	Madrid (m)
bdh0000023	a4644230	40,4199001	-3,68861675	9.78 m		https://pixabay.com/es/m	http://bdh-rd.bne.es/pdf.raw?qu	Puerta de Alcalá	Madrid (m)
bdh0000028	a4645138	40,417417	-3,7125206	12.71 m		https://es.wikipedia.org/v	http://bdh-rd.bne.es/pdf.raw?qu	Plaza de Oriente (Madrid)	Madrid
bdh0000028	a4645199	40,4144273	-3,69106561	2.42 m	Iglesia de los	https://es.wikipedia.org/v	http://bdh-rd.bne.es/pdf.raw?qu	Los jerónimos	Madrid (m)
bdh0000028	a4645803	40,4111679	-3,70876685	19.18 m	La fachada fu		http://bdh-rd.bne.es/pdf.raw?qu	Latina (Madrid, Distrito)	Madrid (m)
bdh0000028	a4646003	40,4241636	-3,7074995	7.7 m			http://bdh-rd.bne.es/pdf.raw?qu	Calle de San Bernardo (Madrid)	Madrid (m)
bdh0000025	a4653640	40,4160814	-3,6964944	18.73 m		https://www.flickr.com/pl	http://bdh-rd.bne.es/pdf.raw?qu	Las Cortes (Madrid, Barrio)	Madrid (m)
bdh0000027	a4649807	40,4257583	-3,70078191	2.52 m		https://es.wikipedia.org/v	http://bdh-rd.bne.es/pdf.raw?qu	Calle de Fuencarral (Madrid)	Madrid
bdh0000027	a4650195	40,4242371	-3,69393826	6.42 m			http://bdh-rd.bne.es/pdf.raw?qu	Salesas, Madrid	Madrid (m)
bdh0000025	a4653995	40,4167798	-3,70348692	13.22 m		https://commons.wikimec	http://bdh-rd.bne.es/pdf.raw?qu	Puerta del Sol (Madrid, Plaza)	Madrid (m)
bdh0000025	a4653723	40,4191976	-3,69312286	10.59 m		https://es.wikipedia.org/v	http://bdh-rd.bne.es/pdf.raw?qu	Fuente de la Cibeles (Madrid)	Madrid (m)
bdh0000027	a4871666	40,4081044	-3,68726492	14.78 m			http://bdh-rd.bne.es/pdf.raw?qu	El Retiro (Madrid)	Madrid (m)
bdh0000067	binp0000225	40,414705	-3,6924845	9.36 m			http://bdh-rd.bne.es/pdf.raw?qu	Paseo del Prado (Madrid, Calle)	Madrid (m)
bdh0000023	a4642431	40,4182174	-3,68436813	2.57 m			http://bdh-rd.bne.es/pdf.raw?qu	El Retiro (Madrid)	Madrid (m)
bdh0000027	a4651095	40,4171678	-3,69389534	2.32 m			http://bdh-rd.bne.es/pdf.raw?qu	Paseo del Prado (Madrid, Calle)	Madrid (m)
bdh0000030	Minp000010	40,4137084	-3,72722983	15.54 m			http://bdh-rd.bne.es/pdf.raw?qu	Calle de Doña Urraca (Madrid)	Madrid (m)

And here data integrated in two geovisualization tools[18]:



[18] Tools used: Google Maps and TimeMapper by the Open Knowledge Foundation Labs.

Initial impact and results

ComunidadBNE was presented in February 6th, with an open workshop at the BNE. While the project is aimed at creating a (new) virtual community, it is clear that direct presence and contact is always an invaluable component to help get qualitative feedback and take corrective actions.

The project was very positively received in every possible aspect; all reactions (at the working session, in social media, or from partners in libraries, research or university communities) showed interest and welcomed the initiative as useful and motivating.

The community currently reaches 298 users (220 work as registered users) and 4475 contributions have been carried out. There are a few extremely active and committed contributors and a good number of high achievers, as shown by the ranking table available in the platform ('Colaboradores' tag), which has by the way proved to be an effective tool to acknowledge and foster participation/competition.

Most successful projects (based on both quantitative –number of volunteers and tasks progress– and qualitative input) are those aimed at georeferencing and text transcription. Clearly, progress is slower (though steady) in projects implying more tasks or requiring additional research. A balance is therefore to be found between different levels of tasks complexity/time required, in order to encourage participation and engagement.

The 'emotional' component is important as well: a project such as the identification of people in photographs from the Spanish Civil War is very unlikely to be completed at all, but any chance of a successful outcome is perceived as worth the effort.

Motivations most usually mentioned by contributors move around these ideas:

- ComunidadBNE as a new way of approaching our heritage, 'fascinating' and 'immersive'.
- A new 'bridge' towards society, whereby the National Library actually 'asks for help' from collective knowledge to improve access and enrich its collections.
- A new way of showing the BNE as an 'open institution', to 'anyone' willing to cooperate (not just researchers or specialists…).
- 'Pride' of belonging to the community and contributing to the BNE's knowledge base for the future.

The future of ComunidadBNE

As mentioned before, the project is currently undergoing a second phase of development. New features and more options for tasks presentation will be available along 2019: navigating and selecting specific tasks within a project; being able to extract an element from an image, describe and store it separately for increased granularity and 'analytical enrichment'; marking irregular areas within an image; audio-to-text transcription/correction; integration of TXT or XML files for OCR or OMR correction.

Special effort will be placed on giving more visibility to contributions and resulting data, through files downloading and data visualization tools. And there is also work to be done on intensifying communication with and *within* the community of contributors.

More crowdsourcing projects are already on their way to ComunidadBNE: identifying and describing advertising pieces and other elements from newspapers; transcribing manuscripts; extracting, tagging and georeferencing elements from a map; tracing characters, subjects, places or references through a novel… are some of the proposals received. A number of academic groups have shown interest in using ComunidadBNE as a tool for their cooperative research.

This is a work in progress that has just started. Leading a crowdsourcing project means learning its specific language and how to take the pulse of the community, looking for the perfect balance between different interests, goals and strategies. But from the experience so far, ComunidadBNE has proved to be a powerful, encouraging and very direct tool for social engagement. Furthermore, it is technically sustainable and entirely aligned with the overall digital strategy at the BNE.

There is still a long way ahead before considering crowdsourcing at the BNE as a consistent, across-the-board data source for authorities and bibliographic records at a mass scale. Initiating the path already entails a significant position change: giving up a certain amount of 'control' over our catalogues in favour of enrichment and social participation still means a great deal to libraries and librarians. But today we already have our linked data repositories enriched with external sources (meaning a clear evolution of the concept of 'catalogue', from a tool for finding and discovering resources to an information source itself) while transparency and social impact/awareness become strategic, institutional key values. Crowdsourcing is just an additional step in the same direction.

If we think of a defining feature for this twenty-first century, we could say that it has radically changed the way we (re)create and distribute knowledge: today it is interdisciplinary by definition, adopts technology as a natural means to evolve, acknowledges collaboration and an open, democratic approach as essential.

And here libraries, and crowdsourcing as a tool with such huge potential, (still) have so much to say.

References

Brabham, D. (2008). *Crowdsourcing as a model for problem solving: an introduction and cases*. Convergence, 14(1).

Howe J. (2006): "The rise of crowdsourcing". *Wired* magazine. 14:1-4.