



HAL
open science

Crowdsourcing, computing and deep mapping cultural heritage and transnational bibliographic records ★

Amel Fraisse, Ben W Brumfield, Sara Carlstead Brumfield

► To cite this version:

Amel Fraisse, Ben W Brumfield, Sara Carlstead Brumfield. Crowdsourcing, computing and deep mapping cultural heritage and transnational bibliographic records ★. Proceedings of the 19th The Conference on Information and Research science Connecting to Digital and Library science, Feb 2023, Bari (Italy), Italy. pp.60-67. hal-04058883

HAL Id: hal-04058883

<https://hal.science/hal-04058883>

Submitted on 5 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Crowdsourcing, computing and deep mapping cultural heritage and transnational bibliographic records[★]

Amel Fraisse^{1,*†}, Ben W. Brumfield^{2†} and Sara Carlstead Brumfield^{2†}

¹Univ. Lille, ULR 4073 - GERiiCO, F-59000 Lille, France

²Brumfield Labs, Texas, USA

Abstract

In an increasingly globalized context, multilingualism and multiculturalism have become major preoccupations for Library and Information Science (LIS) which has to be as fair as possible to ensure and sustain knowledge as a driver for development. This paper proposes a new collaborative, interactive and incremental paradigm for cultural heritage and multilingual bibliographic data curation, processing and mapping. Our first experiment was conducted on the multilingual bibliographic records of the world-famous and well-traveled American novel *Adventures of Huckleberry Finn* by the American author Mark Twain.

Keywords

Digital libraries, multilingual bibliographic records, cultural heritage, crowdsourcing

1. Introduction

The ever growing scientific and political interests in making knowledge open, accessible and sustainable has sparked major interest in many parts of the scientific community. Some disciplines have been concerned with problems of knowledge preservation, organization and dissemination for a long time. Library and Information Science (LIS) is such a discipline [1]. As a gateway to knowledge and culture, the field of LIS holds a long history on collecting, storing, organizing, and sharing access to knowledge. To this purpose, Knowledge Organization Systems (KOS), Information Retrieval Systems (IRS) and metadata exchange standards, among others, have been developed to meet the opportunities arising through the development of new technologies. Collections of the world's great libraries have been made available to the public through large-scale digitization. The Online Computer Library Center (OCLC) dedicated to the public purposes of furthering access to the world's information produces and maintains

19th IRCDL (*The Conference on Information and Research science Connecting to Digital and Library science*), February 23–24, 2023, Bari, Italy

*Corresponding author.

†These authors contributed equally.

✉ amel.fraisse@univ-lille.fr (A. Fraisse); benwbrum@gmail.com (B. W. Brumfield); sara@fromthepage.com (S. C. Brumfield)

🌐 <https://pro.univ-lille.fr/amel-fraisse/> (A. Fraisse)

🆔 0000-0002-8693-8862 (A. Fraisse)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

WorldCat [2], the largest Online Public Access Catalog (OPAC) in the world. WorldCat itemizes the collections of 72,000 libraries in 170 countries and territories. Multilingual online digital libraries and archival projects collect documents and make them available to a wide audience: the Wikisource project (<https://wikisource.org>), an online digital library of free content textual sources, the Internet Archive project (<https://archive.org>) building a digital library of Internet sites and other cultural artifacts in digital form such as books and audio records, or the Gutenberg project (<https://www.gutenberg.org>) offering over 56,000 free written and audio eBooks and especially older works for which copyright has expired in more than 50 under-resourced languages. Those ongoing projects have made and continue to make significant progress in the preservation of knowledge and language diversity.

2. The role of library and information science in building a global, shared knowledge community

More than a century ago, Paul Otlet, the pioneer of Documentation Studies, envisioned a universal compilation of knowledge and the technology to make it globally available. He wrote numerous essays on how to collect and organize the world's knowledge [1, 3]. The ever growing number of digital documents and scientific and political interests in making them openly available all over the world has led to the creation of new digital collections in a broad range of fields and languages. Several Registries of Open Access Repositories (ROARs) hosted by national and international organizations and universities, have been developed. For example, The Library of Congress has digitized approximately 164 million items in virtually all formats, languages, subjects, and periods. These collections are broad in scope, including research materials in more than 470 languages and multiple media. The Europeana collection, launched in 2008 and funded by the European Commission, contains over fifteen million digitized paintings, drawings, maps, photos, books, newspapers, letters, diaries, etc., from fifteen hundred institutions. However, the language barrier is a key issue that Knowledge Organization Systems (KOS) have to address as described in [4, 5, 6]. Indeed, over time, the gap between languages of dominant nations or civilizations and other languages has been growing. Although KOS include knowledge encoded in under-resourced languages, their use and exploration are still limited.

Preserving knowledge diversity and ensuring the right of all people to access knowledge in their mother tongue is the main goal of the Information for All Programme (IFAP) created by UNESCO. Several research work have called for cultural and linguistic diversity as described in [7, 8, 9, 10, 11].

3. Crowdsourcing as a means of decentering institutional authority and expanding knowledge diversity

For documentary edition projects, many take a hybrid approach utilizing crowdsourced transcription for the initial transcription of documents followed by editing and annotation by professional researchers and staff. The Civil War and Reconstruction Governors of Mississippi project works with volunteers to transcribe documents and then research assistants identify

entities – people, places, military units – to index and categories the documents in the published digital documentary edition. The resulting annotated, indexed edition is then published at the scholarly publication site CWRGM.org. Similarly, researchers at Deakin Univeristy transcribed, indexed and annotated the *Howitt and Fison* linguistic surveys, using both academics and members of the communities surveyed in 19th-century Australia. Members of the communities in the anthropological surveys help transcribe, but also identify material in the texts as they interact with the stories and language of their predecessors recorded by colonial anthropologists¹.

Another–more bibliographic–approach to indexing was undertaken by the Texas Commission on Libraries and Archives, who transcribed a handwritten index to their historic Court of Appeals records. This isn't the same kind of record as a diary or a letter–it's not really useful to transcribe this as if it were prose – you need to encode the appellant, appellee, book and page numbers as structured data. The staff configured the project for spreadsheet transcription, and had users encode the contents as if they were typing in excel. The results were exported from FromThePage as a CSV file, and their Preservica export was able to create an interactive lookup table as a finding aid, feeding the CSV file to a javascript library and webpage embedded within Preservica [12].

3.1. Co-creating and crowdsourcing knowledge of folklife and music traditions through the Library of Congress

Traditions of collaborative knowledge creation in cultural heritage are perhaps rarer than they should be, but there are precedents in this sector as well. The twentieth-century folklorist Alan Lomax devoted his life to recording, celebrating, and promoting folk artists and tradition bearers in America, the Caribbean, and Europe. He conducted extensive fieldwork trips during which he produced audio recordings and extensive notes about the people he met, and their traditional arts. His goal was to demonstrate the value of traditional arts, and challenge what he saw as a hegemonic media and cultural system in America and Europe which failed to make room for cultural differences and killed off diversity. As argued in [13], Lomax was critical of “a centralized mediascape through which was broadcast an industrial American mono-culture”. “Too few transmitters and too many receivers” was his central complaint. He was frustrated with the myopic unilateralism of corporate programming, which he saw operating through an “over-powerful, over-rich, over-reaching” communication system. His answer to this was what he termed “cultural equity”: the right for folk communities- what he called “little bubbles of song and delight and ways of life and cookery,” encompassing “hundreds of thousands of these little generators of the original” - to have their voices heard and their traditions represented.” Lomax ultimately re-corded over 1000 cultural groups, and hundreds of under-represented languages. He established the Association for Cultural Equity to advocate for folk artists, and donated his field notebooks, recordings, letters, and other papers to the Library of Congress where he helped to establish the American Folklife Center (AFC). In 2015, the AFC digitized Lomax's papers and made them available online. In 2019, AFC partnered with a new crowdsourcing effort called *By the People* at the Library of Congress, to crowdsource the transcription, review, and tagging of these papers. By the People's goals are to engage a diverse volunteer base with cultural

¹<https://howittandfison.org/about>

heritage preserved at the Library of Congress; to generate transcriptions that will improve online search at the document level, and to provide transcriptions that can be read by screen readers, in order to assist people with visual or cognitive impairments, and those who can't read original handwriting. By the People launched in October 2018 and to date volunteers have transcribed over 100,000 pages from a variety of collections including the papers of Rosa Parks, Walt Whitman, President Abraham Lincoln, and leading suffragists such as Susan B. Anthony and Mary Church Terrell. Volunteers are encouraged through the site itself, emails, in-person events, and social media to explore the documents, ask questions, speak with one another, and Library employees about their findings, struggles, joys, and what they're learning. Their knowledge is taken back into the Library website in the form of transcriptions and enhanced metadata. By the People is a natural extension of Alan Lomax's efforts to build "two-way bridges" and "two-way inter-communication systems" for traditions presented in any medium" as described in [14]. Documents in "The Man Who Recorded the World: On the Road with Alan Lomax" By the People transcription Campaign include materials in Haitian Creole, and dialects of Swedish, Polish, Danish, Hungarian, and other languages spoken by nineteenth- and twentieth-century migrants to the American Midwest, which volunteers transcribe in the original language. In addition to reaching out to over 30,000 registered volunteers to encourage them to participate in the project, AFC folklorists reached out to several descendants of the tradition bearers whom Lomax originally recorded to encourage them to contribute to By the People, and bring their knowledge to bear in this next phase of folklife preservation and exploration.

3.2. The FromThePage platform as a gateway to our cultural heritage

FromThePage is a collaborative textual editing platform released under an open-source license. Its main use is for crowdsourced manuscript transcription by libraries and archives, but it has also been used for researchers preparing digital scholarly editions and for students in classrooms. The system presents an facsimile image to a user alongside a text editing area, saving version history, mark-up, and annotations each the user saves the page.

Since scholars are working with textual material from around the world, multi-lingual support is essential. However, the needs of multi-lingual support rapidly expand the demands on the features of the software. The first non-English projects hosted on the project—A Nahuatl/English edition of the *Codex Aubin* and an annotated Old French/English edition of the *Assizes de Jerusalem*—required the addition of a translation capability to the tool (in addition to existing transcription/OCR text correction functionality), as well as parallel text export in TEI-XML. Subsequent projects in Arabic and Urdu required ISO-639-3 specification of documents in order to present correct justification for Right-to-Left scripts.

Multilingual support requires more than support for editing multilingual texts, however. Volunteers often need "permission" to contribute: they need confidence that their contributions will be welcomed, that their work will be of adequate quality, and that they themselves are "good enough". The software interface language and the language of communication around the project may give them that permission or dissuade them from participating.

Linguists at the University of Texas-Austin crowdsourced transcription of the *Kathryn Josseland Mixtec Language Surveys*, an unpublished collection of field notes. During the project,

one volunteer left a few, short comments in English. When they were invited to communicate in Spanish, their contributions increased by several times. Since this volunteer was a native Mixtec speaker, their contributions went beyond transcription to commentary and annotation on the contents of the surveys[15].

The language of software itself is important for public participation. To better serve communities in Latin America, University of Texas-Austin Libraries won a National Endowment for the Humanities-Office of Digital Humanites grant to fund the translation of the FromThePage interface into Spanish and Portuguese. This enables not only transcription of Spanish and Portuguese texts, but also indigenous-language texts like the colonial documents written in Nahuatl at the Royal Archives of Cholula [16].

4. Our model for expanding and sustaining multilingual bibliographic records

4.1. Crowdsourcing and collaborating for expanding multilingual and transnational bibliographic records

Digital libraries projects have taken up the role of data curation facing a range of highly challenging issues considering the diversity of knowledge encoded in different languages and in particular those encoded in under-resourced ones. Unlike the existing curation model where knowledge is collected only by professional librarians or researchers, we extended the proposed data curation model proposed in [17] to the the collection and the incremental enrichment of multilingual bibliographic records. We conducted a first experiment of this model on Mark Twain texts [18]. Mark Twain's books are some of the most well-travelled texts on the planet. According to the UNESCO Index Translationum², the American writer is ranked 15th in the Top-20 of the most translated authors worldwide. His works have been translated into many languages [19] including under-resourced languages. The novel *Adventures of Huckleberry Finn* is one of the most commonly translated of his books.

Due to the significant number of existing translations and the growing number of digital versions made available online, the crowdsourcing allowed us [17] to gather data that would have otherwise been beyond our reach [20]. Crowdsourcing helped reduce the amount of time spent on the task, increase the variety and the range of the data covered (such as identifying translations which are not indexed in public databases). The parameterization of the crowdsourcing experiment was as follows: as we are looking for translations over the world, we have not limited the geographic location of the contributors. Each task consisted of a set of nine questions (i.e., units in the crowdsourcing terminology). First, we asked people to use search engines or online catalogs to look for existing translations in their native language. Then, we asked them if they could find the translator's name, the first year of publication, the publishing house, the URL of the cover, the bibliographic record if it is available, the list of subjects that could be used to index the translation, and available public digital versions.

²<http://www.unesco.org/xtrans/>

4.2. Interactive and incremental deep mapping of multilingual bibliographic records

Unlike existing knowledge sharing models used by most digital libraries and collections, we propose a new interactive model allowing end-users and volunteer scholars to search, contribute and share their knowledge about an original work through an interactive and online global knowledge map (Figure 1) called Deep Maps in [21].



Figure 1: The global knowledge Map representing existing translations and bibliographic records of *Adventures of Huckleberry Finn* worldwide.

The global knowledge map displays all multilingual bibliographic records about all existing translations of a given original work. Each bibliographic record is represented by a node on the world map (Figure 2), which could be considered “completed” when all required knowledge (translator name, title, editor name, list of subjects, link to related digital catalog) is provided and “partially completed” when it lacks some features. Nodes are updated incrementally by end-users and scholars through the map.

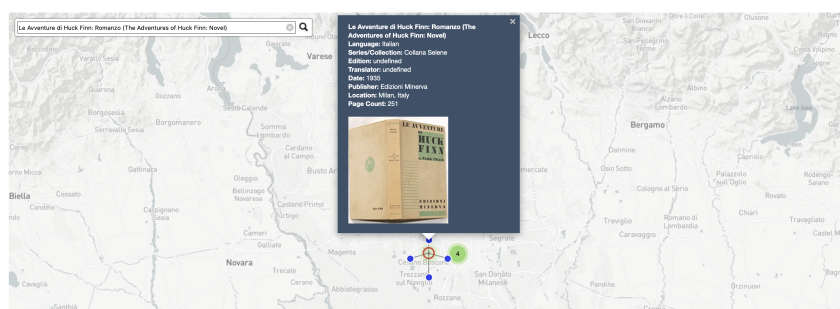


Figure 2: The global knowledge Map representing existing Italian translations of *Adventures of Huckleberry Finn*. In this map, the bubble over Italy is highlighted, displaying the bibliographic record related to the Italian translation.

5. Conclusion

Digital libraries are facing a range of highly challenging issues considering the diversity of knowledge encoded in different languages and in particular those encoded in vulnerable and under-resourced ones. In this paper we described and explored a new paradigm that permits different types of contributors, including volunteers as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines to take part in the data curation and sharing process in an efficient and dynamic way. We explored examples of modern online crowdsourcing, as well as some of the historic attitudes within cultural heritage institutions that have led to or stood in contrast to ideas of co-production or collaboration between institutional gate-keepers and patrons of diverse cultural backgrounds. Crowdsourcing has huge potential to expand the representation of vulnerable languages and cultural practices within the cultural heritage record, and to radically expand the base of people who contribute to the knowledge that is preserved and treated as authoritative by cultural heritage organizations, academia, and other domains.

Acknowledgments

This research work was conducted within the framework of the ROSETTA project funded by the France-Stanford Center For Interdisciplinary Studies in Stanford, USA.

References

- [1] P. Otlet, *Traité de Documentation: le livre sur le livre: théorie et pratique*, Bruxelles: Mundaneum, 1934.
- [2] Ligaya, *Worldcat: Citation tools and features*, *Public Services Quarterly* 6 (2010) 362–363. doi:10.1080/15228951003772488.
- [3] P. Otlet, *Monde: essai d'universalisme: connaissance du monde, sentiment du monde, action organisée et plan du monde*, 1935. URL: <http://dx.doi.org/1854/8321>.
- [4] M. Hudon, *Multilingual thesaurus construction-integrating the views of different cultures in one gateway to knowledge and concepts*, In *Information Services Use* 17 (1997) 11–123.
- [5] M. Hudon, *Accessing documents and information in a world without frontiers*, *The Indexer* 21 (1999).
- [6] A.-H. Barát, *Knowledge organization in the cross-cultural and multicultural society*, In *Advances in Knowledge Organization* 11 (2008) 91–97.
- [7] A. Melissa, J. Tennis, D. Martínez-Ávila, J. A. C. Guima-rães, J.-E. Mai, O. Olesen-Bagneux, L. Skouvig, *Global - local knowledge organization: Contexts and questions*, in: *Proceedings of the Association for Information Science and Technology*, 2006, pp. 1–4.
- [8] B. Clare, *Ethical decision-making for knowledge representation and organization systems for global use*, *Journal of the American Society for Information Science and Technology* 56 (2005) 903–912. doi:<https://doi.org/10.1002/asi.20184>.
- [9] I. Dahlberg, *Ethics and knowledge organization: In memory of dr. s.r. rangathan in his centenary year*, *International Classification* 19 (1992).

- [10] M. López-Huertas, The integration of culture in knowledge organization systems, in: Knowledge Organization for a Sustainable World: Challenges and Perspectives for Cultural, Scientific, and Technological Sharing in a Connected Society, Proceedings of the Fourteenth International ISKO Conference, 2015.
- [11] W. M. E. Hadi, Cultural interoperability and knowledge organization systems, in: Knowledge Organization for a Sustainable World: Challenges and Perspectives for Cultural, Scientific, and Technological Sharing in a Connected Society, Proceedings of the Fourteenth International ISKO Conference, 2016.
- [12] S. Blickhan, B. Brumfield, A. Guzman, V. V. Hynning, The crowdsourcing brick wall: Barriers to data integration and reuse, 2022.
- [13] H. Todd, A. Peart, N. Salsburg, Alan lomax and the grass roots idea, *Chicago Review* 60/61 (2017) 37–45.
- [14] R. baron, All power to the periphery” the public folklore thought of alan lomax, *Journal of Folklore Research* 49 (2012).
- [15] K. Josserand, Transcriptions of mixtec language surveys, 2020. URL: <https://ailla.utexas.org/islandora/object/ailla:271571>.
- [16] S. C. Brumfield, B. Brumfield, Lessons from 5 years of indigenous language transcription projects, 2021.
- [17] A. Fraisse, Z. Zhang, A. Zhai, R. Jenn, S. F. Fishkin, P. Zweigenbaum, L. Favier, W. M. E. Hadi, A sustainable and open access knowledge organization model to preserve cultural heritage and language diversity, *Information* 10 (2019). doi:10.1145/1219092.1219093.
- [18] A. Fraisse, R. Jenn, Q.-T. Tran, Crowdsourcing model for multilingual corpus and knowledge construction: The case of transnational mark twain, *ZIN. Issues in Information Science. Information Studies* 56 (2018).
- [19] R. M. Rodney, *Mark Twain International: A Bibliography and Interpretation of his Worldwide Popularity*, Greenwood Press, Westport, CT, 1982.
- [20] A. Zhai, Z. Zhang, A. Fraisse, R. Jenn, S. F. Fishkin, TI-explorer: A digital humanities tool for mapping and analyzing translated literature, in: Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Barcelona, Spain, 2020.
- [21] S. F. Fishkin, Deep maps: A brief for digital palimpsest mapping projects (dpmps) or ‘deep maps.’, *Journal of Transnational American Studies* 3 (2011). URL: <http://escholarship.org/uc/item/92v100t0>.