Amel Fraisse – Univ. Lille, EA 4073 - GERiiCO, France  
Samantha Blickhan – The Adler Planetarium, Chicago, USA  
Victoria Van Hyning – Library of Congress, Washington, USA  

Towards an Open, Inclusive and Sustainable Knowledge Organization Models

Abstract:  
In an increasingly globalized context, multilingualism and multiculturalism have become major preoccupations for Knowledge Organization (KO) which have to be as fair as possible to ensure and sustain knowledge organization as a driver for development. Indeed, over time, the gap between languages of dominant nations or civilizations and other languages has been growing. In this research, we describe, evaluate and present first results of our sustainable and open access knowledge organization model. The model is based on a paradigms that permit 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 knowledge organization process in an efficient and dynamic way. Recent experiments with this model have been conducted on transnational literary texts as well as in the arena of crowdsourcing cultural heritage knowledge and collections enrichment.

1.0 Introduction

The impact of the digital revolution on the preservation, organization, and sharing of human knowledge encoded by languages constitutes an extraordinarily rich phenomenon, characterized by both productive opportunities as well as obstacles and threats. In the first instance, digital has created tremendous opportunities in terms of accessing knowledge. New technologies also constitute a step forward in terms of public inclusion and awareness. In fact, the general public can be included and integrated, thanks to social networks and collaborative platforms, to provide mass dissemination of human knowledge. Nevertheless, there are numerous barriers that prevent sustained knowledge diversity as described by Hudon (1997), Beghtol (2002), Fraisse et al. (2019), and Barát (2008). Language is the most important barrier; as language diversity is decreasing, the preservation and transmission of such knowledge is at risk. 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 dissemination for a long time. Library and Information Science (LIS) is such a discipline. As a gateway to knowledge and culture, the field of LIS holds a long history on collecting, storing, organizing, and sharing access to knowledge as described by the pioneer of Documentation Studies Paul Otlet (1934). 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 purpose of furthering access to the world's information, produces and maintains WorldCat, the largest online public access catalog (OPAC) in the world. WorldCat itemizes the collections of 72,000 libraries in

© Ergon - ein Verlag in der Nomos Verlagsgesellschaft
170 countries and territories. Multilingual online digital libraries and archival projects collect documents and make them available to a wide audience.

2.0 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 (Otlet 1934). 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 by Hudon (1997; 1998) and Agnes Hajdu Barat (2008). 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 is still limited.

3.0 Current situation

3.1 Crowdsourcing as a means of decentering institutional authority and expanding the representation of different languages and cultures

Since the year 2000 online crowdsourcing projects have proliferated in science, humanities, and cultural heritage fields. Hundreds of cultural heritage institutions have embarked on these projects, many of which explicitly invite people from diverse walks of life to transcribe, annotate or highlight text, speech, typed or handwritten documents. A wide spectrum of languages, historical periods, materials, and geographic areas are represented by these projects and the people who participate in them as described by Van Hyning (2019) and Ridge (2014). As more transcriptions and tags become available in different languages, cultural heritage institutions better represent the peoples and cultures in their collections, and the patrons and communities they serve. By inviting volunteers in to the process of transcribing, translating and tagging, institutions have the opportunity to co-create new knowledge, and make new discovery pathways through collections. These relationships can ultimately decenter traditional power dynamics and concepts of authority in knowledge systems, often for the better, though as Eveleigh (2014) demonstrates, the breakdown of barriers between professional practice and the knowledge of external participants is not inevitable with all crowdsourcing projects, but

1 https://www.loc.gov
2 https://www.europeana.eu
rather requires careful project design and strategies of volunteer engagement. Examples of under-represented language crowdsourcing projects include the City Archive of Leuven³ project to transcribe more than 950,000 Dutch-language register pages from the Leuven court of Aldermen during the years 1362 to 1795; the Ancient Lives project, which launched on the Zooniverse.org platform in 2011, described by Williams et al. (2014), which asked online volunteers to transcribe fragments of ancient Greek texts from papyri fragments; and the Rediscovering Indigenous Languages project ⁴ crowdsourced the transcription of historic word lists, records and other documents relating to indigenous Australian languages. Many of these communities are self-sustaining, self-organizing, and productive of new knowledge, as well as decentralized—participants can translate and describe code and systems, but they can also contribute new functionality to them. This type of community co-creation might serve as a fruitful model for cultural heritage crowdsourcing, in which authority and the creation of descriptive records is still often overwhelmingly concentrated within institutions, rather than shared with the communities that originate cultural artefacts, texts, music, dance, and other outputs.

3.2 Increasing demand and need for global knowledge sharing and access

According to the Sapient Globalization Report there are over 6,700 living languages in the world; the fifteen most popular languages are spoken by 49.5% of the world’s population, while the other 51.5% of the world’s population speak 6,600 languages. Yet, only about 6% of the world’s population speak English. Of the world’s 6000+ languages only a small fraction, a dozen or so, currently enjoy the benefits of modern information technologies and knowledge organization systems. A larger but still modest number, close to a hundred, have the so-called Basic LAnguage Resource Kit (BLARK): monolingual and bilingual corpora, machine readable dictionaries, terminologies, thesauri, ontologies and the like as described by Steven Krauwer (2003) and Antti Arppe et al. (2016). 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 by e.g. Adler et al. (2016), Beghtol (2005), Dahlberg (1992), López-Huertas (2016), and Mustafa El Hadi (2015). In a previous research work Beghtol (1986; 2001) introduce the concept of cultural warrant. Fisher Fishkin (2011) introduced and described a new model for data curation and sharing by inviting colleagues around the world to collaborate on Digital Palimpsest Mapping Projects (DPMPs), or “Deep Maps”. Deep Maps, curated collaboratively by scholars in multiple locations, would put multilingual digital archives around the globe in conversation with one another, using maps as the gateway.

---

³ https://www.itineranova.be/in/home
⁴ https://transcripts.sl.nsw.gov.au
4.0 The open, inclusive and sustainable knowledge organization models

4.1 Basic principles

4.1.1 From closed, discontinuous and out of context to open, continuous and in context knowledge organization models

Our solution aims to move from a closed, discontinuous, and out of context to open, continuous, and in context knowledge organization model. The basic concept is based upon the software localization paradigm proposed by Fraisse (2010) and Fraisse et al. (2009) and promoting the right of all people to use software in their mother tongue. It consists of renouncing the idea of perfect and complete knowledge and publishing partial knowledge with variable quality, which will be improved incrementally during the use of the knowledge organization system. Therefore, the knowledge organization process will be ongoing and improve continuously. The new process permits the incremental augmentation of both quality and quantity. The best known example of this is the Wikipedia community, in which knowledge is added and improved continuously by contributors.

4.1.2 From exclusive, unilateral and unsustainable to inclusive, collaborative and sustainable

Although, current knowledge organization models seems impossible for most languages and even less so for endangered ones, both for reasons of cost, and quite often a scarcity or even lack of expert in these languages. Our solution aims at involving non-experts such as volunteer contributors and especially end-users. These groups have the capacity to participate effectively, since they have a better knowledge of the target language (generally their native language) as well as the context of knowledge being processed.

4.2 The ROSETTA knowledge organization model

An early implementation of the open, inclusive and sustainable knowledge organization model described above was implemented and experimented under the international research project ROSETTA\(^5\) funded by the France-Stanford center for interdisciplinary studies. The main goal of this project consists of defining a knowledge organization model for translated literary texts as well as related scientific documents to these translations. As described by Fraisse et al. (2019), the proposed model is open, inclusive and sustainable. It is based on contributions of end-users as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines. It consists in collecting knowledge about all worldwide translations of one original work and sharing that data through a digital and interactive global knowledge map. The proposed sustainable model allows different types of volunteers and contributors to take a part in the knowledge organization process in an efficient, dynamic and symbiotic way: while using the knowledge map, volunteers and contributors who know the local culture and language can participate by adding missing information about a given translation of an original work. Volunteers and contributors could be scholars or simply citizens interested in preserving knowledge diversity. We define global

knowledge $GK$ about an original work $ow$ as a set of knowledge $K$ about different translations $t$ of $ow$:

$$GK_{ow} = \{K_{ow}^{t_1}, K_{ow}^{t_2}, \ldots, K_{ow}^{t_n}\}$$

where a knowledge $K_{ow}^{t_i}$ about a given translation $t_i$ is a set of key properties as:

$$K_{ow}^{t_i} = \{\text{title}, \text{target language}, \text{translator name}, \text{publication date}, \text{full text}\}$$

Related knowledge was organized collaboratively and in context by scholars through an interactive and online global knowledge map. As described in Figure 1, the map displays all knowledge about all existing documents related to a given literary work. Each document is represented by a node on the world map, which could be considered as “completed” when all required knowledge is provided and “partially completed” when it lacks some knowledge. Nodes are updated incrementally by transnational end-users and scholars through the map. Indeed, during the map exploration, the end-user could edit any node to add missing knowledge.

![Figure 1: The global knowledge map representing existing translations of Adventures of Huckleberry Finn. The bubble over Brazil is highlighted, displaying the relevant information for the Portuguese translations from Brazil.](image)

### 4.3 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 Harvey et al. (2017) argue, Lomax was critical of “a centralized mediascape through which was broadcast an industrial American monoculture”. “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 recorded 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 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 by Baron (2012). 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.

4.4 Crowdsourcing multilingual knowledge: The Zooniverse platform

One current example of Knowledge Organization principles applied through crowdsourcing is Scribes of the Cairo Geniza7, a collaboration between Zooniverse8, the University of Pennsylvania Libraries, and more than half a dozen research institutions who have provided digital geniza images for the project. The Zooniverse is the largest platform in the world for online crowdsourced research, with more than 250 projects launched since its inception in 2009. As of writing, the Zooniverse has more

6 https://crowd.loc.gov
7 https://www.scribesofthecairogeniza.org
8 https://www.zooniverse.org
9 https://www.scribesofthecairogeniza.org/about#provenance
than 1.9 million registered volunteers, who have collectively produced over 460 million classifications on crowdsourcing projects from a variety of disciplines, including astronomy, biology, ecology, climate science, history, and social science as explained in Blickhan et al. (2019). In 2015, Zooniverse launched the Project Builder\(^{10}\), a tool which allows anyone to create and run their own crowdsourcing project, hosted on Zooniverse, free of charge. The platform is maintained by teams based at the University of Oxford (Oxford, UK), the Adler Planetarium (Chicago, IL), and the University of Minnesota Twin Cities (Minneapolis, MN). While the majority of users thus far have been from English-speaking countries, the Zooniverse community is international. To reflect the global userbase, the Zooniverse team created a translation interface, which allows projects to be translated either by project team members, or by volunteers who want to help make a project available for a specific community of speakers. Along with the relatively recent option of a multilingual interface, the Zooniverse has featured multilingual project content since its early days. Ancient Lives\(^{11}\), which launched in 2011, invited volunteers to transcribe fragments of the Oxyrhynchus papyri. In order to open up participation to members of the public who were not fluent in Ancient Greek, the team created a clickable keyboards that volunteers could use to transcribe fragments through character matching. The Scribes of the Cairo Geniza project launched a transcription interface in 2019 with a similar type of assistive keyboard, as well as a multilingual user interface in Arabic, English, and Hebrew. In Scribes of the Cairo Geniza, volunteers are asked to help sort and transcribe fragments of the Cairo Geniza, a corpus of discarded fragments of pre-modern manuscripts discovered in the Ben Ezra synagogue in Fustat (now known as Cairo). The project is broken down into a series of workflows. The Sorting workflow asks volunteers to classify fragments as being written in either Arabic script, Hebrew script, or both. Based on the script type identified, volunteers are then asked to identify specific visual features like page layout, evidence of binding, etc., which can be added to each fragment’s record and assist in the identification process. Once a fragment has been sorted, it is sent to one of four transcription workflows: Easy or Difficult Arabic, or Easy/Difficult Hebrew. Separating workflows based on task type allows volunteers to choose how they wish to contribute based on their comfort level with the different tasks. For example, volunteers who are unable to read Hebrew or Arabic script are able to participate in the Sorting workflow, which offers an introductory tutorial as well as resources on identifying the differences between Hebrew and Arabic script. Volunteers who are fluent in Hebrew and/or Arabic, or who are able to confidently read either script, may choose to participate in the various transcription workflows. The transcription workflows feature clickable keyboards, which function as an additional linguistic and paleographic resource for transcribers who may need additional visual cues to aid in the transcription process.

5.0 Conclusion

Knowledge Organization is 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 an open, inclusive and sustainable knowledge model that permit different types

\(^{10}\) https://www.zooniverse.org/lab
\(^{11}\) https://www.ancientlives.org
of contributors, including volunteers as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines to take part in the knowledge organization 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.

References


Samantha Blickhan, Coleman Krawczyk, Daniel Hanson, Amy Boyer, Andrea Simenstad, Victoria Hying, and Victoria van Hying. 2019. “Individual vs. Collaborative Methods of Crowdsourced Transcription.” Journal of Data Mining and Digital Humanities: hal-02280013v2. https://hal.archives-ouvertes.fr/hal-02280013v2/document


Hudon, Michèle. 1997. “Multilingual Thesaurus Construction-Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts”. In: Information Services and Use 17: 11–123.


Williams, Alex C., John F. Wallin, Haoyu Yu, Marco Perale, Hyrum D. Carroll, Anne-Francoise Lamblin, Lucy Fortson, Dirk Obbink, Chris J.Lintott, and James H. Brusuelas. 2014. “A Computational Pipeline For Crowdsourced Transcriptions Of Ancient Greek Papyrus Fragments”. In 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, 100-105. doi: 10.1109/BigData.2014.7004460.