Multilingual Metadata:

A Brief Overview of Problems and Solutions

Laurel Schwaebe

Department of Library and Information Science, University of Denver

LIS 4010: Organization of Information

Professor Krystyna K. Matusiak

March 19th, 2022

As international information sharing grows with the internet and improved communication channels, user ability to search for multilingual materials is becoming increasingly important. Leaving the metadata for these materials in their original language or transcribed does not allow users to find materials, so another solution must be found. Both machine translation and human translation offer significant hurdles, leading us to look to the future for true multilingual accessibility.

Introduction

With the expansion of the internet and international communication, the issue of multilingual information becomes more and more pressing. Worthwhile intellectual and creative materials are hardly confined to one language, and failing to present materials in accessible metadata results in intellectual isolationism. As Niininen et al. (2016) notes, "Language plays a key role in participating in the global community. Through multilingual and open linked metadata, information can be located and retrieved not only across different collection providers, but also across languages" (p. 452) In order for users to access the greater information world, metadata must be available in other languages with accurate translations and means to find multilingual resources. There are a number of ways to address this problem "Multilingual searching of metadata records requires that either the metadata are provided in multiple languages that users are able to search or user query is translated into language of the metadata," explains Matusiak et al. (2015). Here, we will primarily address the issue as it relates to English speakers finding non-English materials in three key forms of translation.

Example: The Three Body Problem by Liu Cixin[1]

Because no person can read every language found across the globe, translation of some sort will always be necessary. However, language is not so simple that one can simply translate each character individually and expect to understand the author's full meaning. Liu Cixin's (刘慈欣) popular novel, translated as "*The Three Body Problem*" in English, provides an example of the importance of accurate multilingual metadata. This rare Chinese science fiction novel has

[1] Liu is Cixin's family name, listed first here in alignment with Chinese names. Ken Liu shares the same family name as Cixin, but is listed in English order as he is an American.

been translated into dozens of languages and has gained popularity around the world. The English translation by Ken Liu, no relation to Cixin, touched the minds of such celebrities as Barack Obama and George R. R. Martin.

Non-Translation or Transliteration

Metadata that is imported with minimal treatment, either transferred in its original language or transliterated into latin-script characters, is fairly common with non-English materials in English speaking countries. Naturally, the original language text is ultimately the best interpretation of various metadata terms, but it highly limits searchability, readability, context, and other information that users need to utilize these resources.

A quick review of *the Three Body Problem* on WorldCat brings up an entry for "*三体/San ti*," the original Chinese title for the book. While the author's transliterated name will connect a user with the English version of the book, as well as other translations and the rest of the series, one might not necessarily connect *三体/San ti* with *the Three Body Problem*. To start with, the name is not translated into the same name used in the English version. Additionally, the transliteration of *三体*, "*San t*i," does not feature the tones necessary for determining the characters, which may lead to confusion in searching for the words used in the title. Without the tones, "San ti" can be interpreted as "sǎntí,[2]" which may mean "carry umbrella," or as "sàntī," which means "free kick." However, even the inclusion of pinyin tones does not necessarily confirm accurate translation, as many Chinese characters use the same pinyin and have different meanings. Additionally, the series title, *地球往事* (translated as *the*

[2] It is worth noting that even Calibri font, the default font for Google Documents, lacks certain glyphs used in pinyin, such as ǎ and ī are shown in an alternate font.

Remembrance of Earth's Past in English), is listed as "*Zhongguo ke huan ji shi cong shu*" on WorldCat, which neither immediately brings up the series, nor give the reader any indication of what the series actually is, but instead provides a number of other Chinese books with transliterated names. Further research reveals that this series title is actually the name of a Chinese science fiction magazine in which the first of *the Remembrance of Earth's Past* series was published and is not related to the individual entry for the Chinese book.

Furthermore, in the wrong font or without the proper language drivers installed, the title becomes "□□" which is entirely unsearchable to speakers of any language. Many machines or engines struggle to recognize unique letters, not limited to Chinese, the machine will replace unfamiliar glyphs with placeholders (Vanderbilt et al., 2010). We see this with the boxes that substitute for Chinese characters, and also sometimes with glyphs like the letter "ø" in Nordic languages. Such substitutions can entirely distort the text's meaning, or even dramatically change it, such as in the example shown in figure 1.

In other cases, the characters may be replaced with codes made up of a series of numbers that indicate what the character was. Vanderbilt et al. (2010) notes "Chinese data are not in ASCII (American Standard Code for Information Interchange) character format and when written into Metacat are changed to html Unicode format" (p. 189). For example, "中文" (zhōngwén), meaning "Chinese language,"was translated to "中&#25991". While this may provide a more accurate indication of what the character was than the box glyph, it is certainly not readable by the average Chinese speaker, let alone someone unfamiliar with the language. Without translation, a reader looking for the original text may be left wondering how to search in Chinese characters or searching in pinyin.

AI Translation

A Machine translation service "automates the process of language translation by analyzing and understanding information in one language and expressing it in another language." (Matusiak et al., 2015, p. 4) Services like this include Google Translate, and can be useful in navigating non-English websites. However, everyone who has used these services has a story of translation gone wrong. While translating in one direction seems easy enough, reviewing what the words actually mean, especially when brought back to English, demonstrates the imperfection of both the English language and in AI translation. In the words of Ludwig Wittgenstein, "You approach from one side and know your way about; you approach the same place from another side and no longer know your way about." (Niininen et al., 2016, p. 453) In early 2014, a meme video challenge dared content creators to Google Translate popular songs into other languages, and then back into English, often with humorous results. One such video by Twisted Translations sent the lyrics to the hit Disney song "Let It Go" from the animated film "Frozen" through transitions into Chinese, French, Macedonian, Polish, Creole, Tamil, and then back into English with baffling results (figure 2).

In this instance, we can see that the translations have become confused at some point because we can read both the translation and the original material. However, if we do not know the original language, we cannot understand what nuances are lost, what words are changed in machine translation.

Human Translation

Naturally, one might feel that the best thing would be human translation. Indeed, human translators can capture the implicit meaning behind words and create more rounded,

authentic translations. In our example, the original title reads as "三体," or literally "Three Body," was adjusted to "The Three Body Problem" in English to clarify its reference to the physics conundrum. The "体" (tǐ) character sounds similar to "题" (tí), which means "problem" or "question," making the title a pun in Chinese. While Ken Liu's translation may not be the most accurate to the individual characters in the title, the meaning is well preserved. There is little doubt that translations created by professionals with the proper research and context for the materials are the most accurate and are the most useful to users.

Yet, human translation is not as widespread as one might expect. "In contrast to automated methods, it renders more accurate metadata and captures the cultural nuances but is resource and cost intensive," explains Matusiak et al. (p.14, 2015). Translation website the Translation Company lists translation at between $0.09 and $0.40 per word in the United States, and can cost up to 100% more per word for specialized technical language. Assuming the institution does not have the budget for massive translation projects, it could rely on its native speaking staff for translation. However, multilingual staff is not a given, and even were such a person within the institution, this project would consume much of their time. Additionally, no language is perfectly aligned with another, and as such, translations will always be imperfect. Regarding the Tse-Tsung Chow Collection of Chinese Scrolls and Fan Paintings, Matusiak et al. (2015) said: "Without additional contextual information, the meaning and cultural significance of digitized objects may not be readily apparent not only to English language users but also to Chinese language speakers who may not be familiar with traditional Chinese styles and scripts" (p.14). So in addition to being able to translate, staff members must also then have significant knowledge about the materials or be able to research the materials

thoroughly. Furthermore, metadata terms are not consistent throughout the world, which may lead to misalignment in metadata, the terms for both languages had to be interpreted and aligned. This process “relies on expertise of the translator for the semantic accuracy of the translation.” (Vanderbilt et al., 2010, p. 193) Matusiak et al. (2015) demonstrates a multilingual pay of metadata terms, a process which might be simplified with metadata schema registries that provide schemas mapped to one another. As Sugimoto (2014) statesm “proper localisation of metadata labels, and management and maintenance of URIs of metadata terms are the fundamental requirements for metadata interoperability across communities and over time” (p. 67).

The Future

Machine translation may not currently be a viable solution to multilingual metadata, but it does present a path forward. It is impractical to expect that libraries could handle the level of translation needed for their resources, both from a financial and labor standpoint, and from an accuracy standpoint. “We are… unaware of any software that is more than experimental that allows users to see the same concepts through the lenses of different languages,” meaning that accurate translation requires expertise and context that may not always be available to an individual translator (Vanderbilt et al., 2010). Because of this, AI is likely to provide the ultimate translation solution. Already, we have seen vast improvements in systems like Google Translate, with translations between linguistically similar languages approaching full accuracy. While languages like Chinese are not likely to reach this level in the coming years, one might expect that technology will improve to this point in the future with the help of AI content training and community teaching.

In the meantime, AI can learn from existing community-sourced metadata. Kawakami et al. (2020) reported keywords supplied by multilingual volunteers for images of paper cranes. While not all of the keywords were necessarily accurate to the images, they provided cultural insight into interpretations of the images. Focusing the keywords by certain categories (appearance, name, and function) reduced biases in keywords selected and "help workers without full knowledge of the objects shown to focus on categories where they can contribute" (p. 203). This method has its flaws, but pairing it with advancing machine translation may provide the balance between ease and accuracy needed until the AI can produce accurate results on its own.

Conclusion

Ultimately, the search for solutions to multilingual metadata ends in disappointment, but with hope for the future. No current option provides the balance needed between cost and time efficiency and accuracy. Machine translation can quickly provide metadata translations, but may be inaccurate or misleading. Human translation can offer the best translations, complete with context, but take time, money, and extensive knowledge of each resource. While there are significant flaws with both of these options, they are better than simply leaving the metadata untranslated or transliterated. While no solution is forthcoming at this time, we can see a path to quick and accurate machine translation in the future, ideally supplemented with community input to provide necessary context.

Sources

Kawakami, M., Sakaguchi, T., Shirai, T., Matsubara, M., Yoshino, T., & Morishima, A. (2020). Analysis of crowdsourced multilingual keywords in the futaba digital archive: Lessons learned for better metadata collection. In *Digital Libraries at Times of Massive Societal Transition* (pp. 196–204). Springer International Publishing. https://doi.org/10.1007/978-3-030-64452-9_17

"Let It Go" from Frozen according to Google Translate (PARODY) [Video file]. (2014, February 10). In *YouTube*. Retrieved March 3, 2022, from https://www.youtube.com/watch?v=2bVAoVlFYf0

Liu C. (2008). *The Three Body Problem*. Chongqing Publishing House.

Matusiak, K., Meng, L., Barczyk, E., & Shih, C.J. (2015). Multilingual metadata for cultural heritage materials: The case of the Tse-Tsung Chow Collection of Chinese scrolls and fan paintings. *The Electronic Library, 33*(1), 136 - 151.

Niininen, Nykyri, S., & Suominen, O. (2017). The future of metadata: open, linked, and multilingual – the YSO case. *Journal of Documentation, 73*(3), 451–465. https://doi.org/10.1108/JD-06-2016-0084

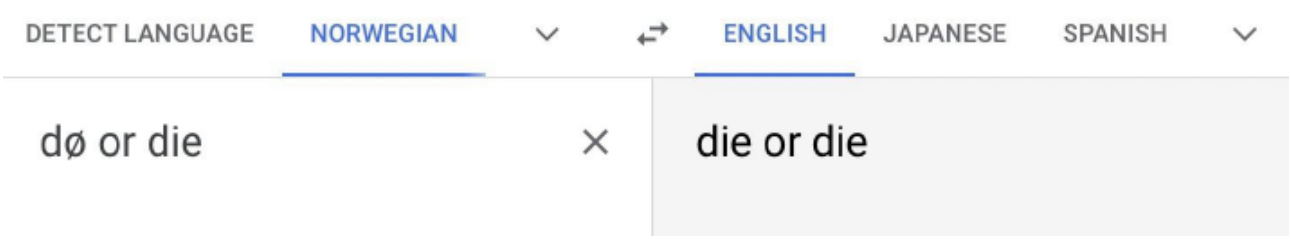
Sugimoto, Shigeo (2014) Digital archives and metadata as critical infrastructure to keep community memory safe for the future – lessons from Japanese activities, *Archives and Manuscripts, 42*:1, 61-72, DOI: 10.1080/01576895.2014.893833

The Translation Company (2022, January 10). *5 facts you should know to buy translation*. The Translation Company. Retrieved March 3, 2022, from https://thetranslationcompany.com/resources/translation-infographics.htm

Vanderbilt, Blankman, D., Guo, X., He, H., Lin, C.-C., Lu, S.-S., Ogawa, A., Ó Tuama, É., Schentz, H., & Su, W. (2010). A multilingual metadata catalog for the ILTER: Issues and approaches. *Ecological Informatics, 5*(3), 187–193. https://doi.org/10.1016/j.ecoinf.2010.02.002

三体 / *San ti*. (2008). Retrieved March 3, 2022, from 三体 / San ti. (2018). Retrieved March 3, 2022, from https://www.worldcat.org/title/san-ti/oclc/1102268445&referer=brief_results#relatedsubjects.

Figure 1.



Screenshot taken from Google Translate on 3/3/2022.

Figure 2.

English	Google Translation
The snow glows white on the mountain tonight *Not a footprint to be seen* *A kingdom of isolation,* *And it looks like I'm the Queen.* *The wind is howling like this swirling storm inside* *Couldn't keep it in, heaven knows I tried* *Don't let them in, don't let them see* *Be the good girl you always have to be* *Conceal, don't feel, don't let them know* *Well, now they know* *Let it go, let it go* *Can't hold it back anymore* *Let it go, let it go* *Turn away and slam the door*	*Lit white snow on the mountain tonight* *No visible legs,* *Discrimination law* *is probably the Queen.* *Rotating the wind is howling storm,* *They cannot do that, God knows I've tried.* *Do not let them, do not let them see* *It is always a good girl.* *Hide, do not feel, do not know* *Well now you know* *Give up, give up* *You can not do it back in* *Give up, give up* *Tune in, and slam the door*

Lyrics transcribed from *Let It Go" from Frozen according to Google Translate (PARODY)* video on 3/3/2022.

Figure 3.

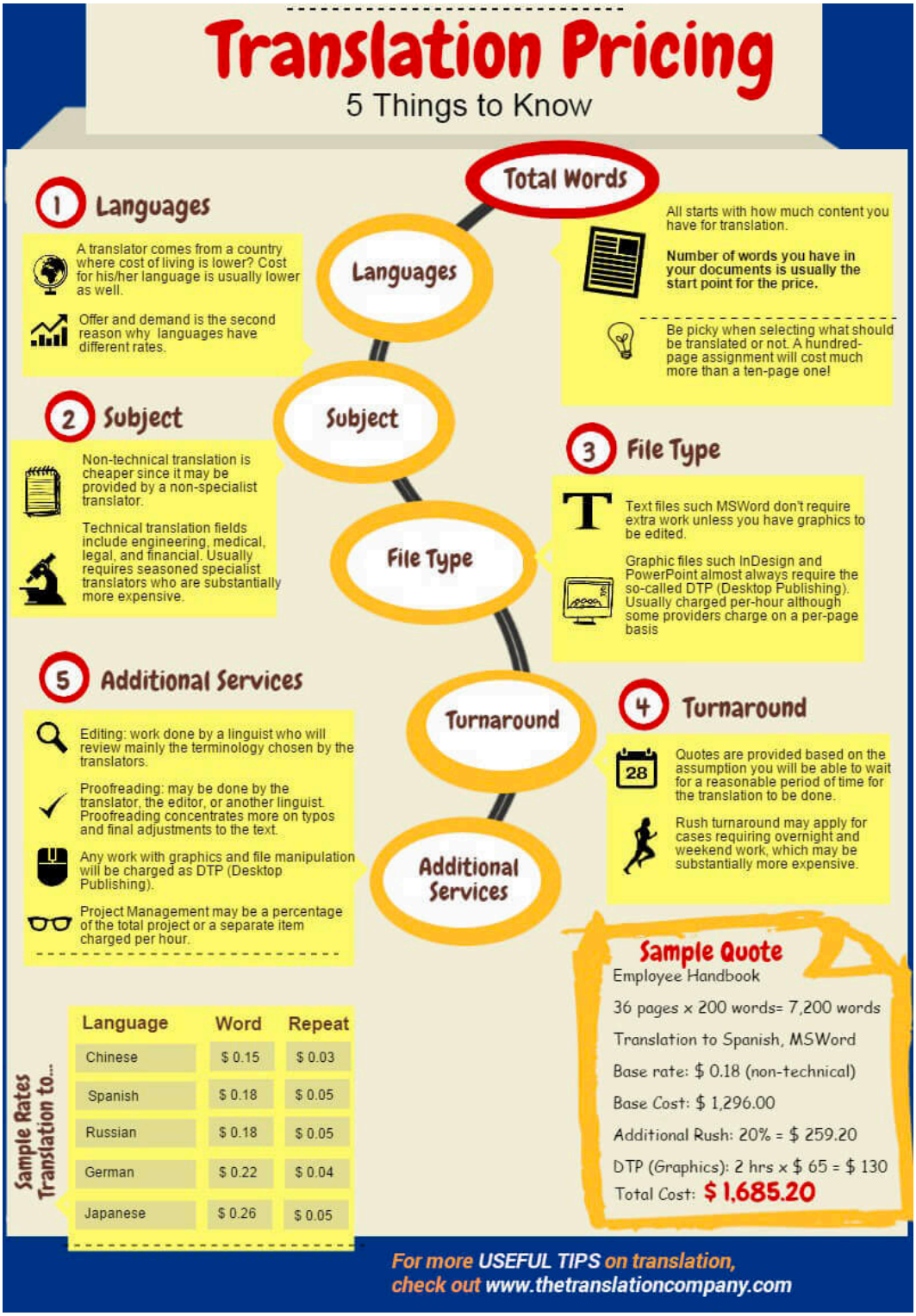


Image taken from the Translation Company on 3/3/2022.

Figure 4.

	Field name	DC map	Data type	Large	Search	Hide	Required	Vocab
1	Title/標題	Title	Text	No	Yes	No	Yes	No
2	Artist	Creator	Text	No	Yes	No	No	Yes
3	作者	Creator	Text	No	Yes	No	No	Yes
4	Artist Bio	Description	Text	Yes	No	No	No	No
5	作者生平	Description	Text	Yes	No	No	No	No
6	Date of Creation	Date-Created	Text	No	Yes	No	No	No
7	創作日期	Date-Created	Text	No	Yes	No	No	No
8	Period	Date	Text	No	Yes	No	No	No
9	年代	Date	Text	No	Yes	No	No	No
10	Collector	Contributors	Text	No	Yes	No	No	Yes
11	收藏者	Contributors	Text	No	Yes	No	No	Yes
12	Description	Description	Text	Yes	No	No	No	No
13	描述	Description	Text	Yes	No	No	No	No
14	Main Text	Description	Text	Yes	Yes	No	No	No
15	釋文	Description	Text	Yes	Yes	No	No	No
16	Other Text	Description	Text	Yes	Yes	No	No	No
17	題跋與款識	Description	Text	Yes	Yes	No	No	No
18	Seal Content	Description	Text	No	Yes	No	No	No
19	印記	Description	Text	No	Yes	No	No	No
20	Language	Language	Text	No	Yes	No	No	No
21	語言	Language	Text	No	Yes	No	No	No
22	Subject AAT	Subject	Text	No	Yes	No	No	Yes
23	主題 AAT-Taiwan	Subject	Text	No	Yes	No	No	Yes

Table taken from Matusiak et al.