Dealing with multilingualism and non-English content in open repositories: Challenges and perspectives

Christophe Dony – https://orcid.org/0000-0001-7421-5993
Iryna Kuchma – https://orcid.org/0000-0002-2064-3439
Milica Ševkušić – https://orcid.org/0000-0002-2888-6611

Note that this is a preprint version of an article that has been accepted for publication in a special issue on 'Multilingual Publishing & Scholarship' in *The Journal of Electronic Publishing* (Michigan Publishing).

Abstract

Several organizations and initiatives have recently called for more support of multilingualism in research to promote epistemic plurality and raise awareness of the adverse effects of an anglocentric research ecosystem. But this support for and practice of multilingualism and linguistic diversity cannot happen in a digital or technological vacuum. Open repositories can play an important role in ensuring that research infrastructures have the ability to implement and promote multilingualism at scale in an Open Science environment. This implementation, however, is complex and does not come without its own theoretical and technical challenges. One of these challenges is to recognize that the implementation of multilingualism in open repositories can hardly be dissociated from wider concerns of discoverability, research assessment practices, and the anglocentric nature of digital infrastructures and metadata standards or protocols. Drawing on the COAR (Coalition of Open Access Repositories) recommendations report produced by the COAR Task Force on Supporting Multilingualism and non-English Content in Repositories, this article presents and critically examines how and why three particular recommendations of this document are particularly well suited to support a decolonial trajectory for the management of multilingualism in open repositories. More specifically, this article discusses the decolonial aspects and praxis underlying guidelines such as declaring the language(s) of the resource and of its metadata, writing personal name/s using the writing system used in the deposited document while providing a persistent identifier to disambiguate author/s identification and, overlapping with the latter, enabling UTF-8 support so as to promote use of the original alphabet / the writing system whenever possible, without negating the possibility to transliterate metadata by means of recognized standards (e.g. ISO). In so doing, we argue that these recommendations enable a multifaceted technology and politics of recovery that promotes a form of linguistic revitalization and strengthens linguistic diversity.

10.5281/zenodo.11060284

Introduction

Advocacy for multilingualism in research has recently gained momentum, notably thanks to key reform- and policy-oriented texts related to Open Science and research assessment initiatives. The Helsinki Initiative on Multilingualism in Scholarly Communication (Federation Of Finnish Learned Societies et al. 2019), the Call For Action to foster bibliodiversity in scholarly communications (Shearer et al. 2020), the UNESCO recommendation on Open Science (UNESCO 2021), and the Agreement of the Coalition for Advancing Research Assessment (Science Europe and COARA 2022), for example, all support that multilingualism is necessary to help develop and maintain a diverse and qualitative research landscape. All of these texts suggest, albeit differently and in varying degrees, that the hegemonic status of English in research threatens bibliodiversity, hampers research innovation, and limits the development and significance of "locally relevant" research (Federation Of Finnish Learned Societies et al. 2019).

While this defense for a gradual acceptance and improved recognition of multilingualism and non-English content in research is both timely and important for advancing equity, inclusivity, and social engagement in the global research landscape, it cannot happen without an enhanced discoverability capacity and the adoption of a particular knowledge-sharing and archiving practices, let alone thrive in a global knowledge ecosystem that is increasingly digital, connected, and versed in dynamics of interoperability and semantic and linked data. Because of their community-oriented agenda setting and their ability to promote alternative circuits of of publishing and knowledge dissemination (see e.g. Chan et al. 2019; Collyer 2018), open repositories and archives play an important role in defining and framing a knowledge-sharing and archiving praxis that improves the digital curation, management, and discoverability of multilingual or non-English content.

In August 2022, the Confederation of Open Access Repositories (COAR), an international organization aiming to build a "inclusive and trusted global knowledge commons based on a network of open access digital repositories" (COAR Confederation of Open Access Repositories, n.d.), launched a dedicated Task Force to develop and promote good practices for repositories in managing multilingual and non-anglophone content. In October 2023, the Task Force published its recommendations document (COAR Task Force on Supporting Multilingualism and non-English Content in Repositories 2023), which presents a series of guidelines and good practices based on the community input that the Task Force received after a public consultation. Eight recommendations on creating and curating metadata and six recommendations for repository software and platform developers were generated. They focus on declaring the language(s) of the resources and their metadata, using standard language codes, ensuring language specific user interfaces can be used, proper inclusion of personal names, using multilingual keywords, vocabulary and thesauri, and proper management of translated content.

The very nature and scope of these recommendations, just like their implementation, does not come without practical and theoretical challenges, especially as they relate to linguistic marginalization and, more generally, decolonial perspectives on and about multilingualism and archival practices, both of which have a longstanding history with

colonization and nation-building (see e.g. Carbajal 2021; Ghaddar and Caswell 2019; Gramling 2021; Ndhlovu and Makalela 2021; R'boul 2022a; Said 1979; Williams, Deumert, and Milani 2022). For example, digital architecture in general (Kwet 2019) and tools used to build digital archives such as open repositories have often been designed from a Western universalist perspective or unique ontology allegedly usable across different languages and cultures (see Chaka 2022; Filimowicz 2023; Graham and Dittus 2022). In fact, in most cases English is the lingua franca for such systems and tools. This is not without posing serious technical issues in terms of flexibility as it can relate to the co-existence of languages and scripts. In a similar decolonial and postcolonial perspective, the metadata schemes and controlled vocabularies that are used to describe content in digital libraries and open repositories for enhanced discovery and interoperability purposes may not appropriately document Indigenous traditional knowledge and local languages or properly accommodate translated or multilingual content. The integration of multilingual keywords to open repositories, which could allow users to discover scholarly content in multiple languages, represents yet another challenge. For interoperability and discovery purposes, it should ideally be based on mapping strategies of common existing schemes whose ontologies are far from being equally inclusive to various cultural contexts, social groups, and languages (see Drabinski 2013; Howard and Knowlton 2018; Lacey 2018; Vaughan 2018).

In light of these numerous challenges, it is therefore important to further contextualize the COAR recommendations for the management of multilingual and non-English content through a decolonial critical lens, so as to reflect on their potentially decentering effects and inherent limits or tensions. It is, of course, beyond the scope of this paper to discuss all of the recommendations of the COAR document in this perspective. The authors, who participated in the COAR Task Force, therefore identified particular recommendations to critically engage with in this paper from a decolonial perspective in its broadest sense. These recommendations include: declaring the language(s) of the resource and of its metadata, writing personal name/s using the writing system used in the deposited document while providing a persistent identifier enabling unambiguous author/s identification and, overlapping with the latter, enabling UTF-8 support so as to promote use the original alphabet / the writing system whenever possible, without negating the possibility, if necessary, to transliterate metadata by means of recognized standards (e.g. ISO).

The objectives behind our selection of these particular recommendations are manifold. From a visibility, discovery, and evaluation perspective, we argue that these recommendations can help foster a "balanced multilingualism", which considers that all forms and languages needed for all research purposes must be recognized and documented to improve "the monitoring of further globalization of research" and ensure more diversity and equity in processes of research evaluation (Siversten, 2018), while allowing better discoverability beyond default English settings of particular digital systems and architectures. Further, we posit that this balanced multilingualism and the recommendations that it is built on can intimate what Kim Gallon calls "a technology of recovery" in exploring Black Digital Humanities (Gallon 2016), i.e. a conceptual framework which aims at recovering formerly marginalized voices and content through the use of digital platforms and resonates with the decolonial ethics of language reclamation theories and practices that place language revitalization and visibility beyond purely linguistic observations (Leonard 2012; 2017; Grenoble and Whaley 2021; Engman, Hermes, and Schick 2022; Filimowicz 2023). In the context of the management of multilingualism and non-English content in open repositories, it is important to note that this

“technology of recovery”, and its attendant politics, is predicated on the implementation of particular technical or strategic developments that run the risk of reproducing some of the politics of exclusion and marginalization, linguistic or otherwise, that are embedded in the very design and processes of research discovery platforms and archival practices.

To better understand and engage with the tensions underlying these overlapping objectives, it is therefore first useful and necessary to briefly contextualize the politics of multilingualism as they can be envisioned in the context of the geopolitics knowledge dissemination and its attendant digital infrastructures. This is what the section below sets out to do without pretending to be exhaustive.

Literature Review

There is no denying that having a scientific lingua franca such as English “facilitates scientific mobility and [...] collaboration”, just like it facilitates “international scientific communication” and “dissemination” (Steigerwald et al. 2022, 988). However, maintaining and supporting a monolingual research landscape has many disadvantages. It can be detrimental to global evidence synthesis and regional or community-oriented policy-making (Amano et al. 2021; Amano, Berdejo-Espinola, et al. 2023; Angulo et al. 2021; Konno et al. 2020). It can also reinforce the standardization and homogenization of research practices as language is constitutive of how we perceive, explore, describe and analyze the world (cf. Angulo et al. 2021; Hsu 2017; R’boul 2022b). Moreover, it places extra labor efforts and difficulties on non-English researchers, whose lack of language proficiency can lead to various gatekeeping effects - editorial or otherwise (cf. Amano, Ramírez-Castañeda, et al. 2023; Lillis 2010; Uzuner 2008). All in all, limiting the production and dissemination of knowledge to a common language can lead to various types of injustice and a lack of epistemological diversity, for which sociologist Boaventura de Sousa Santos has coined the term “epistemicide” in his decolonial exploration of knowledge theory (see Santos 2011; 2018).

The inherent limits to knowledge diversity in the global research landscape have much to do with the Englishization of research, or rather with the promotion of English as a “standard for research visibility” (R’boul 2022a, 144), which concerns what is regarded, valued, and counted as knowledge or research. University rankings and what they are based on play a crucial role in this matter (cf. Kris and Robertson 2016; Morrisson 2021; St Clair 2021; Stack and Ishikawa 2021). In particular, the two major anglo-centric bibliographic and citations indexes used in many rankings and broader evaluation processes worldwide (Kulczycki 2023), namely the Web of Science and Scopus, downplay the importance of “the contributions of universities beyond the Anglosphere” (St Clair 2021, 133). Both indexes are indeed widely known for privileging anglophone content and journals (Vera-Baceta, Thelwall, and Kousha 2019; Tennant 2020; Khanna et al. 2022; Bardiau and Dony 2024). Vera-Baceta et al.’s study estimated the proportion of English content in Scopus and WoS at 92,64% and 95,37% respectively (Vera-Baceta, Thelwall, and Kousha 2019, 1806). Though the scope of Scopus is admittedly more international (Baas et al. 2020), the selection processes of these indexes is particularly problematic in terms of linguistic diversity as both require that journals’ article titles and abstracts be translated into English (see Clarivate, n.d.; Elsevier, n.d.), thus omitting non-Latin scripts and writing systems.

Researchers and evaluators' heavy reliance on these commercial indexes, and on the metrics and rankings that they are based on (Collyer 2018; Kulczycki 2023; Kris and Robertson 2016; Stack and Ishikawa 2021; Morales et al. 2021), contributes to the subordination of non-English research in discovery tools and research evaluation processes (St Clair 2021; Schmidt 2020; Mamdani 2019), despite the fact that the development of non-English research is still vibrant when looking beyond these indexes. For instance, it is well recorded that research in the social sciences and the humanities (SSH) is often "grounded in specific cultural or geographical areas" and therefore promotes the "the persistence of native languages" rather than solely focusing on English (Giglia 2019, 143). This persistence of native or local languages in SSH research is attributed to various forms of social engagement with cultural and political concerns (Giglia 2019; Kulczycki et al. 2020; Luzón 2019), which can be considered to represent local and alternative hubs of knowledge. This is particularly true in light of the fact that multilingual publishing has also reportedly been presented as "an ongoing practice in many SSH research fields regardless of geographical location, political situation, and/or historical heritage" (Kulczycki et al. 2020, 1371). Moving beyond SSH, looking at multidisciplinary journals lists and digital libraries beyond traditional indexes such as Scopus and the WoS show that the scholarly communications landscape embraces linguistic diversity and multilingual publishing more than is generally assumed. For example, a recent study analyzing the 25,671 active journals employing the open-source publishing platform Open Journal Systems (OJS) only reports a 49,7% proportion of journals using English as a main language of publication (Khanna et al. 2022). Building on this study, Mikael Laakso and Janne Pölönen have attempted to map languages used in a global landscape made of 150,760 scholarly journals and reported a proportion of journals using English only at 47%, with journals using multiple languages at 19% (Laakso and Pölönen 2023). Open Access repositories and digital libraries also ensure the preservation of many non-English and multilingual scholarly content beyond journal articles. As of December 6th, 2023, for example, the Directory of Open Access Books (DOAB) indexed over 76,000 books with more that 30,000 books published in languages other than English (DOAB, n.d.). Similarly, in a recent Report on Repository Survey in Europe, it was shown that a majority of open repositories surveyed "collect content in at least two languages" (Shearer et al. 2023, 28), albeit with "either the main local language being most predominant, or second most predominant after English"(Shearer et al. 2023, 29). As of December 6th, 2023, the recently launched open scholarly communications digital catalog OpenAlex (Priem, Piwowar, and Orr 2022) indexed more 246,800,000 scholarly objects, of which over 74,000,000 are allegedly not in English.

This non exhaustive list of examples attests to the sheer volume of multilingual and non-English content in digital libraries and repositories, most of which are part of a growing and multidimensional "alternative, open discovery infrastructure" that "builds on a network of tens of thousands of libraries, archives, repositories, and aggregators that offer their (meta-)data via an open data interface such as OAI-PMH" or similar metadata protocols (Kraker, Schramm, and Kittel 2021, 5). This growing open discovery infrastructure, with its multidimensionality and decentralized governance, can help us challenge a universalist and English-centered perception of the research ecosystem (cf. Chan et al. 2019), especially as currently defined in North America and Europe through the lens of traditional, yet somewhat outdated, commercial discovery indexes and metrics. Ensuring that research infrastructures have the ability to implement and improve the digital curation and discoverability of multilingual or non-English content *at scale* in an Open Science environment is of crucial importance in this respect. But paradoxically enough, there is very little guidance on how to tackle these

questions and issues at scale, even if some very general guidelines (Diekema 2012; Wu and Chen 2022) exist, just like presentations of community-scaled and -specific initiatives or cataloguing developments related to digital libraries and archives (see, e.g. Concordia, Gradmann, and Siebinga 2010; Stiller et al. 2014; Matusiak et al. 2015; Riva 2022).

General guidelines and recommendations provided by Diekema (2012) and Wu and Chen (2022) identify major challenges and obstacles for the management of multilingualism in digital libraries. Diekema's review primarily offers a presentation of technical challenges as they can relate to technical aspects, including "data management (localization and language processing), representation (dealing with different fonts and character codes), development (creating international software, cross-cultural collaboration), and interoperability (system architecture and data sharing)" (Diekema 2012, 165). Drawing on The World Digital Library and the Digital Library of the Caribbean as case examples of successful multilingual libraries, Wu and Chen (2022) primarily focus on the organizational and operational obstacles needed to sustain multilinguality in these digital environments (Wu and Chen 2022). They emphasize the need for partnership and collaboration as well as fundraising and budgeting capabilities to envision an ongoing and sustainable development and implementation of multilingualism in such environments. In the case examples studied, the authors highlight that fundraising and grants allowed the creation of particular digital library software and application software which helped meet the specific multilingual needs and objectives of these projects. Both studies, however, devote little attention to technical specifications and how they can convey a particular trajectory to multilingualism - ideological or otherwise.

Other studies have shown the importance of translating multilingual metadata schemes and keywords to allow for enhanced discoverability and to improve multilingual retrieval search functions for similar community-scaled or -specific initiatives and projects (see, e.g. Concordia, Gradmann, and Siebinga 2010; Stiller et al. 2014; Matusiak et al. 2015; Riva 2022), while sometimes also pointing to the possibility of crowdsourcing for doing so (Budzise-Weaver, Chen, and Mitchell 2012). Processes of participatory metadata and objects description for scaling up such endeavors have also been recommended (Haberstock 2020). Of particular interest in terms of research discoverability is the recent development of the GoTriple project, a European discovery platform for Social Sciences and Humanities which supports discovery in 12 languages thanks to an advanced approach of metadata enrichment that is based on a "hierarchical" and multilingual thesaurus of "over 3.300 SSH-related concepts in these 12 languages: Croatian, Dutch (partial), English, French, Finnish, German, Greek, Italian, Polish, Portuguese, Spanish, Ukrainian" (GoTriple, n.d.; see also Dumouchel et al. 2020). The structure of this vocabulary, however, can be criticized for perpetuating Western-oriented hierarchical and institutionalized logic of subject descriptions, which in this case heavily draws on the useful, yet in many regards contested, Library of Congress Subject Headings Classification (see Drabinski 2013; Howard and Knowlton 2018; Lacey 2018; Vaughan 2018).

While these works show that solutions retained for implementing multilingualism in digital archives and libraries are very much context-specific and that recommendations can hardly be imagined according to a prescriptivist logic, they mainly focus on providing insights on technical solutions, platform organization, or infrastructure sustainability. As a result, they usually fail to critically engage with how their proposed technical solutions or design enhancements may be at odds with the decentering logic of a decolonial archival praxis, which

“considers how archives emerge through multifaceted global processes and structures, and are embedded within larger discursive formations, in which multiple cultural sites, texts and contexts are active” (Ghaddar and Caswell 2019, 78). This, of course, may be due to the very dynamic and highly political character of multilingualism itself (cf. Ferrante, Bernstein, and Gironzetti 2019; Gramling 2021; Turner 2023), which remains a moving target and therefore requires ongoing re-evaluation (see Makoni, Kaiper-Marquez, and Mokwena 2022; McKinney, Makoe, and Zavala 2023).

Methods

As previously suggested, this article is grounded in several and intertwining lines of inquiry that draw on decolonial studies, Southern theory, scholarly communications, as well as digital and archival studies, all of which are used to critically engage with how the implementation of particular recommendations for the development of multilingualism and non-English content in open repositories can promote a balanced multilingualism and, at the same time, strengthen a technology and politics of recovery for non-English scholarly content in the global research ecosystem.

The central approach of this article is thus qualitative insofar as it aims to shed light on the (de)colonial realities and mechanisms underlying the recommendations presented here. In so doing, the present work can be located in the continuity of a growing body of archival and library-related scholarship that is concerned with diversifying, decentering, and decolonising scholarly communications and research libraries (see, e.g. Crilly 2023; also see Schmidt 2020; 2023).

Because the corpus of recommendations analyzed here is directly drawn from a specific document written by the COAR Task Force on Supporting Multilingualism and non-English Content in Repositories, it is also important to briefly present this Task Force and to draw the contours of its work, especially so given the topics at stake.

To ensure a diversity of perspectives, the Task Force was composed of a multiplicity of stakeholders (repository managers and translators, representatives of aggregating and discovery systems) coming from various countries (Argentina, Belgium, Canada, China, Ecuador, Germany, Japan, Nepal, Mexico, Peru, Serbia, Spain, Türkiye, Ukraine, and the US). The Task Force, drawing on several use cases contributed from different stakeholders communities, identified three key areas to work on: enhancing discoverability of non-English content, curating multilingual content in a repository, and supporting translations. In June 2023, the Task Force released a preliminary set of draft suggestions for community feedback. The ensuing consultation yielded a diversity of perspectives which were examined before being integrated into the final recommendations document.

The consultation also revealed some limitations and challenges. These are prevailingly associated with technical issues – missing or insufficiently developed features in widely used repository software platforms, the lack of inclusive ontologies, and the lack of standards that would address the specific features of indigenous cultures. We address some of these limitations as they apply to the particular recommendations discussed below.

Writing systems and names

Many use cases presented to the COAR working group involved issues concerning the ability to render text in a variety of writing systems without compromising discoverability, including questions revolving around transcription and transliteration practices and the ability to properly render names and other information (e.g. metadata) in non-Roman alphabets. This led the taskforce to develop a set of particular recommendations addressing these overlapping issues or parts thereof. These recommendations read as follows :

- "Enable UTF-8 support in your repository and use the original alphabet / the writing system whenever possible. If it is necessary to transliterate metadata, use recognized standards (e.g. ISO)"
- If the repository software supports multiple interface languages, set up the user interface in the native language(s) of the target group, along with the English option;
- Write personal name/s using the writing system used in the deposited document and provide a persistent identifier enabling unambiguous identification, such as ORCID" (COAR Task Force on Supporting Multilingualism and non-English Content in Repositories 2023).

These seemingly simple recommendations promote a type of balanced multilingualism that enacts what can be described as a technology and politics of recovery insofar as they improve the visibility of formerly marginalized voices and scripts in digital spaces without compromising discoverability, while enabling greater curation accuracy. To better understand how these intertwining issues underlie the above recommendations and how the latter *write back* to various forms of linguistic exclusion, it is useful to account for how technical limitations of text rendering tools in both pre-digital and digital age have historically affected cataloging practices and metadata curation in digital repositories, including the development of transliteration[1] and transcription[2] techniques as a response to these limitations.

The ability to use a language in a digital space is determined by the ability of technology to support the appropriate writing system. The key technology enabling multilingualism in the digital sphere is the Unicode encoding standard (Korpela 2006), which can support 161 scripts ("Supported Scripts," n.d.), and the UTF-8 variant of Unicode is currently the dominant encoding on the Internet, with 98.1% of surveyed websites using it ("Usage Statistics of Character Encodings for Websites," n.d.). However, the early development of digital technologies and the Internet was marked by the domination of the English language and ASCII (American Standard Code for Information Interchange) character set, which is suitable for English but does not contain any characters with accents or diacritics used in French, Scandinavian and Slavic languages, let alone non-Roman characters (Nolan 2006). Digital technologies and devices designed in and for anglophone environments, with interfaces and supporting documentation in English and restrictive licences hindering localization, set a linguistic barrier and led to the exclusion and marginalization of non-English speaking users (Nolan 2006; Souphavanh and Karoonboonyanan 2005; Mikami and Shigeaki 2012; John 2013). This also limited the ability of non-anglophone communities, especially those not using

[1] Representing characters of one alphabet using the characters of another.

[2] Representing the pronunciation of a term in one language using the characters of the writing system of another language ("ISO 5127:2017(En), Information and Documentation — Foundation and Vocabulary," n.d.)

the Latin alphabet, to express themselves and communicate in digital spaces, raising concerns that “digital colonialism” (Kwet 2019; Kupfer and Muyumba 2022) or “computer-mediated colonization” (Ess 2007) would create digitally disadvantaged languages, that is, languages which are inadequately supported by digital tools such as text processing software, keyboards, fonts, web browsers, OCR (optical character recognition) tools, assistive technologies, etc. (Zaugg, Hossain, and Molloy 2022), and can eventually disappear as a result (UNECSO 2015).

The early versions of Unicode appeared in the early 1990s , but it took over a decade before it was implemented in widely used writing tools, cataloging and repository software, and general and scholarly information retrieval systems. Unicode has been presented as a means “to simplify software internationalization” (John 2013, 329; also see Souphavanh and Karoonboonyanan 2005) and is therefore claimed to function as “yet another instance of western cultural imperialism” (John 2013, 330). In the context of digital repositories, however, Unicode can be perceived differently as its wide adoption has functioned as a solid base for more digital inclusion and the recovery of ‘minority’ languages. Unicode indeed took root in free and open-source repository software in the early 2000s[3] thanks to liberal licensing practices (Souphavanh and Karoonboonyanan 2005), which enabled the development of localized and multilingual user interfaces, metadata input using various scripts and, consequently, support for search strings in various languages and scripts.

Despite these technological prerequisites, the analysis of use cases presented to the COAR working group showed that encouragement to use their full potential is needed, hence the support for UTF-8 implementation in the COAR recommendation. Before the advent of Unicode and UTF-8 and their subsequent adoption in digital repositories, temporary fixes and workarounds for encoding and cataloging practices were developed, e.g. replacing non-supported characters with similar ASCII characters or with images, national encodings, etc. (Korpela 2006; Hardie 2007; John 2013). And some of the use cases submitted to the Task Force revealed that some of these techniques tend to persist even after the emergence of technologies that provide more efficient support for multilingualism.

Transliteration and transcription are examples of such a workaround inherited from the pre-digital age. Similarly to translation (Shamma 2018), transliteration and transcription are in many cases associated with cultural hegemony - e.g. Early Modern missionary dictionaries, translations and writings in vernacular languages printed in the Latin alphabet (Burke 2006; Kiaer et al. 2022; Liu 2018), or the use of first Latin and then Cyrillic for Turkic languages in the Soviet Union (Alpatov 2017). In the context of archival and library practices, transliteration and transcription are a staple of cataloging standards which respond to the need for a single authorized form of a personal name.[4] In the Global North, this single form has always been Romanized - see, for instance, the ALA-LC Romanization Tables (Barry, Library of Congress,

[3] Based on the software documentation, it seems that both DSpace and EPrints supported UTF-8 from the outset, though many fixes were required to make it work properly (“DSpace System Documentation: Version History” 2005; “DSpace Character Encoding HOWTO,” n.d.; “Unicode” 2021).

[4] What makes matters even more complicated is that, in some languages, the transcription of personal names, proper nouns and even loanwords is enshrined in the orthography and legislation (Hardie 2007; Klyshinsky, Maximov, and Yolkeen 2008; Naumova 2014).

and American Library Association 1997) or ISO standards relating to the transliteration of different writing systems ("ISO - 01.140.10 - Writing and Transliteration," n.d.).

This particular approach is arguably associated with the technical limitations of text rendering tools and information retrieval systems which did not support multiple alphabets at the time when the standards were defined. Paradoxically, however, this method has failed to ensure the desired unification due to the multiplicity of systems used for transliteration and transcription. Although disputed as inaccurate, expensive and inefficient (Aissing 1995; Dagher and Soufi 2021), the practice of transliterating and transcribing names in repositories is still widespread for several reasons including: heavy reliance on transcription- and transliteration-friendly metadata standards such Datacite's (DataCite, n.d.), the persistence of traditional cataloging infrastructures and workflows, the substantial body of legacy metadata (inherited from the pre-digital age), and the fear that aggregators and retrieval systems will not be able to process non-Roman characters appropriately. Moreover, it is usually feared that target audiences in the Global North will not be able to decipher names in non-Roman alphabets. However, most search engines can process various languages and scripts, even if risks exist that the ranking algorithms may favor anglophone content and the Roman alphabet in search results (Rovira, Codina, and Lopezosa 2021). Finally, some writing systems are still not encoded in Unicode ("Unsupported Scripts," n.d.) and multilingual support varies across infrastructures and tools. However, it is noteworthy that transliterated and particularly transcribed forms of names can make information retrieval more difficult because users are not necessarily aware of the transformations to which names are subjected in the curation process, some of which can also lead to information loss (Borgman 1997; Monyela 2021).

The COAR recommendation concerned with names tackles these intertwining issues at the technical level, releasing metadata curators from the burden of seeking for a single optimal authorized name form by either transliterating/transcribing it or by recovering its original spelling from the information provided in the publication. According to this recommendation, names in the repository metadata should accurately capture the spelling provided in publications, while disambiguation is to be ensured via unique personal persistent identifiers included in the metadata and linked to external services that store and maintain them (e.g. ORCID, ISNI, VIAF). The advantage of this approach is at least twofold. First, it ensures that content is discoverable regardless of the spelling in the search string. Second, it allows names to be displayed and processed as authors have chosen to render them in the publication.

To complement this approach, the second part of the UTF-8-related recommendation is also grounded in practical logic. It advises to use the writing system of the resource whenever possible, even for metadata that cannot be associated with a persistent identifier (e.g. tites) so as to promote digital inclusion and improve curation accuracy, while at the same avoiding issues of comprehension that could surface from transliteration and transcription standards. Finally, the discoverability of content in non-Roman writing systems can additionally be supported by providing keywords in multiple languages, a possibility that is also addressed in the COAR recommendations.

Declaring languages

Language declaration presents another challenge for multilingual scholarly content discovery. This is because if "the language of a scholarly resource is not labeled properly it will not be correctly indexed by discovery services. That is because indexing involves text analysis practices such as stemming, lemmatization (grouping together the inflected forms of a word so they can be analyzed as a single item), and the appropriate treatment of stop-words. All of these text analysis techniques are very language specific." (COAR Task Force on Supporting Multilingualism and non-English Content in Repositories 2023). Content aggregators and discovery systems therefore need to know the languages of full text documents they index, so they can assist users in finding content in their preferred languages. Repositories and other content management systems therefore need to provide this information by declaring the languages of their resources at the item level and in the resource descriptions (i.e. metadata) to help information seekers and content aggregators, indexers, and discovery services to correctly identify the language of the full text, process the items accordingly and offer better multilingual retrieval. By the same token, declaring the language(s) of a document and that of its metadata can help aggregators and discovery services display languages as filters or in search elements. In turn, this displaying of language(s) as constitutive of a resource can potentially pave the way for newer forms of research monitoring and evaluation, thereby actively contributing to the implementation of a more "balanced multilingualism" (Siversten, 2018). This is also why if the resource (e.g. an edited volume) has important sections of the text in different languages, the language metadata must be repeated to mention each language.

The technical implications underlying the recommendation to declare the language(s) of a resource should not be overlooked as language is both a descriptive and technical characteristic of the resource and a significant property for long-term preservation that impacts rendering, behavior, interpretation and accessibility of digital objects, together with other technical features such as file format, compression algorithm, software version, resolution and color space. This is why the COAR report recommends that language is encoded as a significant property using particular metadata standards often employed together for preserving and managing digital objects, namely PREMIS (Preservation Metadata: Implementation Strategies) and METS (Metadata Encoding and Transmission Standard). As noted in the report, "METS is primarily focused on encoding descriptive, administrative, and structural metadata, providing a framework for organizing and linking various types of metadata within a structured XML document. PREMIS, on the other hand, focuses on documenting the actions, events, and processes involved in the long-term preservation of digital objects. METS can serve as a container for various metadata, including PREMIS metadata, allowing for the integration of preservation-specific information within the broader context of digital object organization and description" (COAR Task Force on Supporting Multilingualism and non-English Content in Repositories 2023). This is why the report recommends that language is encoded as a significant property using PREMIS and considered to be technical metadata, significant for preservation. Language can also be embedded into the METS as technical metadata for text documents. In addition, language information, if considered as a descriptive characteristic of the intellectual content, can be embedded into the METS as descriptive metadata.

The language of the metadata elements - resource descriptions, should be specified as well for the same reasons as outlined above. Regardless of the fact that English is mainly assumed to be the standard for metadata fields, this content should also be exposed with a reference to the language used. It is worth doing it at the repository level as most content aggregators can not infer language from the content of the metadata. Some aggregators, e.g. OpenAIRE supports the language tag and conducts metadata checks for languages in subjects, titles and descriptions. However, there is no exposure of the language of metadata in the exchange protocol used by content aggregators and repositories - Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). As a result, the report invites repository software developers to consider this in future versions of their platforms.

Various approaches are used by repository administrators and managers to declare the language, depending on the capacity of the repository software to handle this information. Some repository software, e.g. WEKO developed by the National Institute of Informatics, Japan, and based on INVENIO by CERN, allows adding a language attribute to any metadata as long as it is allowed in the supported JPCOAR (Japan Consortium for Open Access Repositories) metadata schema. In other cases, new versions of repository software enable language declaration, e.g. new metadata enhancements on Open Science Framework (OSF) for all OSF Projects, Registrations, and Preprints now include the language of materials. Some other repository softwares should be customized, e.g. EPrints repository software can be extended to declare language information at the item or file level but this is not in place on EPrints by default. Similarly EPrints XML export plugins, embedded metadata and OAI-PMH interface code could be extended to define xml:lang attributes but it does not do this by default.

To ensure interoperability between different systems, and hence a better visibility and recognition of language attributes in a variety of platforms that make up the multidimensional alternative, open discovery infrastructure of the research ecosystem, the language metadata must be encoded using a standardized nomenclature to classify languages - the ISO-639 language code is the form of a two- or three-letter, such as ‘en’ or ‘eng’ for English. However, while the ISO-639 use is straightforward for well-known and widely spread languages (in January 2023 it included codes for over 7900 languages), lesser-known languages and regional varieties or historical stages of languages may not be sufficiently represented in ISO 639. To solve this issue, the language code can be followed by optional sub-tags refining or narrowing the range of the encoded language in the following form: language-extlang-script-region-variantextension-privateuse with the “x” private-use sub-tag for the identification of language variations (as described in Gillis-Webber and Tittel 2020, 639). The COAR report includes a decision tree on how to determine a language tag.

Repository software provides multiple ways to implement these recommendations. For example, in DSpace 7, the value-pairs set for languages can include any languages and language identifiers. By default, DSpace provides ten languages value-pairs: English (United States) (en_US), English (en), Spanish (es), German (de), French (fr), Italian (it), Japanese (ja) and Chinese (zh), Portuguese (pt), Turkish (tr). However, it is fully customizable and can include three letter identifiers. During content submissions, language values are displayed as a dropdown list while in the metadata editing mode, language is a free text field. There are also solutions to fix language code inconsistencies in repository platforms.

The implementation of these recommendations can be read as the decolonial action of reclaiming and reassigning value to non-English content via technical processes of localization and multilingual support in digital platforms, which were previously overlooked in the context of anglocentric research and the allegedly universal character of digital infrastructures. It does require extra time and labor, but we believe that the benefits overweight the costs insofar as they help to improve a diversity of various cultural contexts, social groups, and languages, thereby, enabling epistemological diversity (see Santos 2011; 2018) and ensuring that more diversity and equity in research evaluation can be achieved through further fostering of a "balanced multilingualism" (Siversten, 2018).

Conclusion and next steps

The promotion and advancement of multilingualism in research can hardly be decoupled from wider concerns of discoverability, research assessment and monitoring practices, and the anglocentrism of digital infrastructures and metadata standards or protocols. This is why engaging with these intertwining issues and debates is necessary in crafting and providing recommendations for the management of multilingual content in digital spaces. To put it differently, there can only be *ongoing trajectories* for the promotion and advancement of multilingualism in research and scholarly communications. In this article, we have presented and discussed how and why particular recommendations elaborated by a dedicated COAR Task Force instill a decolonial trajectory for the management of multilingual and non-English language content in open repositories. The decolonial aspects of this trajectory can be seen in how the curation practices and technical guidelines embedded in these recommendations enable a multifaceted technology and politics of recovery that promotes a form of linguistic revitalization (see e.g. O'Grady 2018; Grenoble and Whaley 2021; Olko and Sallabank 2021) as well as strengthens linguistic diversity and, eventually, epistemic plurality.

Processes akin to linguistic revitalization and other practices enabling the disruption of the existing anglocentric research ecosystem obviously go well beyond open repositories and the particular recommendations discussed in this article. The COAR recommendations document, for example, also provides guidelines for the management of translated content and advises to "include keywords in many languages" and to "use multilingual vocabularies and thesauri if possible" (COAR Task Force on Supporting Multilingualism and non-English Content in Repositories 2023) to further enhance the discoverability and visibility of non-English content. Next to open repositories, aggregators and discovery platforms should also develop or finetune guidelines and mechanisms to better process and display language-related metadata. Similarly, preprint servers, publishers, and other digital infrastructures archiving or producing scholarly content should also strive to better manage and document multilingualism, including translations. Finally, institutions should also develop strategies and commitments to advance and promote multilingualism in research, including mechanisms to improve its recognition or integration in research assessment .

In the long run, only a wider adoption of practices and recommendations espousing a decolonial trajectory of multilingualism in research will offer possibilities to potentially *decenter* English and recalibrate the volume of non-English content in an otherwise anglocentric research system and its equally anglocentric digital architecture. And because undoing and

unlearning are staple practices of decolonial thinking and praxis (see e.g. Torres 2017; Montgomery and Trahar 2023; Schmidt 2023), the development of standards and recommendations for the support and management of multilingualism in research should remain a moving target, which should notably strive to involve so far marginalized or excluded groups in this process.

Works Cited

Aissing, Alena L. 1995. “Cyrillic Transliteration and Its Users.” *College & Research Libraries* 56 (3). https://doi.org/10.5860/crl_56_03_207.

Alpatov, Vladimir Mikhajlovich. 2017. “Scripts and Politics in the USSR.” *Studi Slavistici*, November, 9–19. https://doi.org/10.13128/Studi_Slavis-21936.

Amano, Tatsuya, Violeta Berdejo-Espinola, Munemitsu Akasaka, Milton A. U. de Andrade Junior, Ndayizeye Blaise, Julia Checco, F. Gözde Çilingir, et al. 2023. “The Role of Non-English-Language Science in Informing National Biodiversity Assessments.” *Nature Sustainability* 6 (7): 845–54. https://doi.org/10.1038/s41893-023-01087-8.

Amano, Tatsuya, Violeta Berdejo-Espinola, Alec P. Christie, Kate Willott, Munemitsu Akasaka, András Báldi, Anna Berthinussen, et al. 2021. “Tapping into Non-English-Language Science for the Conservation of Global Biodiversity.” *PLOS Biology* 19 (10): e3001296. https://doi.org/10.1371/journal.pbio.3001296.

Amano, Tatsuya, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, et al. 2023. “The Manifold Costs of Being a Non-Native English Speaker in Science.” *PLOS Biology* 21 (7): e3002184. https://doi.org/10.1371/journal.pbio.3002184.

Angulo, Elena, Christophe Diagne, Liliana Ballesteros-Mejia, Tasnime Adamjy, Danish A. Ahmed, Evgeny Akulov, Achyut K. Banerjee, et al. 2021. “Non-English Languages Enrich Scientific Knowledge: The Example of Economic Costs of Biological Invasions.” *Science of The Total Environment* 775 (June): 144441. https://doi.org/10.1016/j.scitotenv.2020.144441.

Baas, Jeroen, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. 2020. “Scopus as a Curated, High-Quality Bibliometric Data Source for Academic Research in Quantitative Science Studies.” *Quantitative Science Studies* 1 (1): 377–86. https://doi.org/10.1162/qss_a_00019.

Bardiau, Marjorie, and Christophe Dony. 2024. “Measuring Back: Bibliodiversity and the Journal Impact Factor™ Brand, a Case Study of IF-Journals Included in the 2021 Journal Citations Report™.” *Insights the UKSG Journal* 37: 1. https://doi.org/10.1629/uksg.633.

Barry, Randall K., Library of Congress, and American Library Association, eds. 1997. *ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts*. 1997 ed. Washington: Cataloging Distribution Service, Library of Congress. https://catalog.hathitrust.org/Record/011397188.

Borgman, Christine L. 1997. “Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries: Or How Do We Exchange Data In 400 Languages?” *D-Lib Magazine* 3 (6). https://doi.org/10.1045/june97-borgman.

Budzise-Weaver, Tina, Jiangping Chen, and Mikhaela Mitchell. 2012. “Collaboration and Crowdsourcing: The Cases of Multilingual Digital Libraries.” *The Electronic Library* 30 (2): 220–32. https://doi.org/10.1108/02640471211221340.

Burke, Peter. 2006. “2. The Jesuits and the Art of Translation in Early Modern Europe.” In *2. The Jesuits and the Art of Translation in Early Modern Europe*, 24–32. University of Toronto Press. https://doi.org/10.3138/9781442681552-007.

Carbajal, Itza A. 2021. “Historical Metadata Debt: Confronting Colonial and Racist Legacies Through a Post-Custodial Metadata Praxis.” *Across the Disciplines* 18 (1–2): 91–107.

https://doi.org/10.37514/ATD-J.2021.18.1-2.08.
Chaka, Chaka. 2022. “Digital Marginalization, Data Marginalization, and Algorithmic Exclusions: A Critical Southern Decolonial Approach to Datafication, Algorithms, and Digital Citizenship from the Souths.” *Journal of E-Learning and Knowledge Society* 18 (3): 83–95. https://doi.org/10.20368/1971-8829/1135678.
Chan, Leslie, Angela Okune, Rebecca Hillyer, Denisse Albornoz, and Alejandro Posada, eds. 2019. *Contextualizing Openness: Situating Open Science*. http://ruor.uottawa.ca/handle/10393/39849.
Clarivate. n.d. “Web of Science Journal Evaluation Process and Selection Criteria.” *Clarivate* (blog). Accessed December 5, 2023. https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/web-of-science-core-collection/editorial-selection-process/editorial-selection-process/.
COAR Confederation of Open Access Repositories. n.d. “About COAR.” COAR Coalition of Open Acccess Repositories. Accessed December 1, 2023. https://www.coar-repositories.org/about-coar/.
COAR Task Force on Supporting Multilingualism and non-English Content in Repositories. 2023. “Good Practice Advice for Managing Multilingual and Non-English Language Content in Repositories.” COAR. https://doi.org/10.5281/ZENODO.10053918.
Collyer, Fran M. 2018. “Global Patterns in the Publishing of Academic Knowledge: Global North, Global South.” *Current Sociology* 66 (1): 56–73. https://doi.org/10.1177/0011392116680020.
Concordia, Cesare, Stefan Gradmann, and Sjoerd Siebinga. 2010. “Not Just Another Portal, Not Just Another Digital Library: A Portrait of Europeana as an Application Program Interface.” *IFLA Journal* 36 (1): 61–69. https://doi.org/10.1177/0340035209360764.
Crilly, Jess. 2023. “Diversifying, Decentering and Decolonising Academic Libraries: A Literature Review.” *New Review of Academic Librarianship* 0 (0): 1–41. https://doi.org/10.1080/13614533.2023.2287450.
Dagher, Iman, and Denise Soufi. 2021. “Authority Control of Arabic Personal Names: RDA and Beyond.” *Cataloging & Classification Quarterly* 59 (2–3): 260–80. https://doi.org/10.1080/01639374.2020.1845896.
DataCite. n.d. “2. Creator.” DataCite Metadata Schema 4.5 Documentation. Accessed December 11, 2023. https://datacite-metadata-schema.readthedocs.io/en/4.5_draft/properties/mandatory/property_creator.html.
Diekema, Anne R. 2012. “Multilinguality in the Digital Library: A Review.” *The Electronic Library* 30 (2): 165–81. https://doi.org/10.1108/02640471211221313.
DOAB. n.d. “DOAB - Browsing by Language ‘English.’” https://directory.doabooks.org/browse?type=language&value=English.
Drabinski, Emily. 2013. “Queering the Catalog: Queer Theory and the Politics of Correction.” *The Library Quarterly* 83 (2): 94–111. https://doi.org/10.1086/669547.
“DSpace Character Encoding HOWTO.” n.d. DSpace - LYRASIS Wiki. Accessed December 21, 2023. https://wiki.lyrasis.org/pages/viewpage.action?pageId=22021520#DSpaceCharacterEncodingHOWTO-jira.
“DSpace System Documentation: Version History.” 2005. 2005. https://fenix-ashes.ist.utl.pt/open/trunk/dspace/dspace1.3.2/docs/history.html.
Dumouchel, Suzanne, Emilie Blotière, Gert Breitfuss, Yin Chen, Francesca Di Donato, Maria Eskevich, Paula Forbes, et al. 2020. “GOTRIPLE: A User-Centric Process to Develop a Discovery Platform.” *Information* 11 (12): 563. https://doi.org/10.3390/info11120563.
Elsevier. n.d. “Scopus Content Policy and Selection | Elsevier.” Elsevier. Accessed November 14, 2023. https://www.elsevier.com/products/scopus/content/content-policy-and-selection.
Engman, Mel. M, Mary Rose Hermes, and Anna Schick. 2022. “Co-Conspiring with Land: What Decolonizing with Indigenous Land and Language Have to Teach Us 1.” In *The Routledge Handbook of Language and the Global*

South/s, edited by Sinfree Makoni, Anna Kaiper-Marquez, and Lorato Mokwena. Routledge.
Ess, Charles. 2007. “From Computer-Mediated Colonization to Culturally Aware ICT Usage and Design:” In , edited by Sri Kurniawan and Panayiotis Zaphiris, 178–97. IGI Global. https://doi.org/10.4018/978-1-59904-096-7.ch008.
Federation Of Finnish Learned Societies, The Committee For Public Information, The Finnish Association For Scholarly Publishing, Universities Norway, and European Network For Research Evaluation In The Social Sciences And The Humanities. 2019. “Helsinki Initiative on Multilingualism in Scholarly Communication.” figshare. https://figshare.com/articles/Helsinki_Initiative_on_Multilingualism_in_Scholarly_Communication/7887059.
Ferrante, Laura Di, Katie A. Bernstein, and Elisa Gironzetti. 2019. “Towards Decentering English: Practices and Challenges of a Multilingual Academic Journal.” *Critical Multilingualism Studies* 7 (1): 105–23. https://cms.arizona.edu/index.php/multilingual/article/view/177.
Filimowicz, Michael. 2023. *Decolonizing Data: Algorithms and Society*. London: Routledge.
Gallon, Kim. 2016. “Making a Case for the Black Digital Humanities.” In *Debates in the Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein, 42–49. University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.
Ghaddar, J. J., and Michelle Caswell. 2019. “‘To Go beyond’: Towards a Decolonial Archival Praxis.” *Archival Science* 19 (2): 71–85. https://doi.org/10.1007/s10502-019-09311-1.
Giglia, Elena. 2019. “OPERAS: Bringing the Long Tail of Social Sciences and Humanities into Open Science.” *JLIS.It* 10 (1): 140–56. https://doi.org/10.4403/jlis.it-12523.
Gillis-Webber, Frances, and Sabine Tittel. 2020. “A Framework for Shared Agreement of Language Tags beyond ISO 639.” In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 3333–39. Marseille: European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.408.pdf.
GoTriple. n.d. “GoTriple - About TRIPLE.” GoTriple - The Innovative Discovery Service. Accessed December 7, 2023. https://gotriple.eu/about.
Graham, Mark, and Martin Dittus. 2022. *Geographies of Digital Exclusion: Data and Inequality*. Pluto Press. https://doi.org/10.2307/j.ctv272452n.
Gramling, David. 2021. *The Invention of Multilingualism*. Key Topics in Applied Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108780667.
Grenoble, Lenore A., and Lindsay J. Whaley. 2021. “Toward a New Conceptualisation of Language Revitalisation.” *Journal of Multilingual and Multicultural Development* 42 (10): 911–26. https://doi.org/10.1080/01434632.2020.1827645.
Haberstock, Lauren. 2020. “Participatory Description: Decolonizing Descriptive Methodologies in Archives.” *Archival Science* 20 (2): 125–38. https://doi.org/10.1007/s10502-019-09328-6.
Hardie, Andrew. 2007. “From Legacy Encodings to Unicode: The Graphical and Logical Principles in the Scripts of South Asia.” *Language Resources and Evaluation* 41 (1): 1–25. https://doi.org/10.1007/s10579-006-9003-7.
Howard, Sara A., and Steven A. Knowlton. 2018. “Browsing through Bias: The Library of Congress Classification and Subject Headings for African American Studies and LGBTQIA Studies.” *Library Trends* 67 (1): 74–88. https://doi.org/10.1353/lib.2018.0026.
Hsu, Funie. 2017. “Resisting the Coloniality of English: A Research Review of Strategies.” *The Catesol Journal* 29 (1): 111–32. https://files.eric.ed.gov/fulltext/EJ1144339.pdf.
“ISO - 01.140.10 - Writing and Transliteration.” n.d. International Organization for Standardization. Accessed December 11, 2023. https://www.iso.org/ics/01.140.10/x/.
“ISO 5127:2017(En), Information and Documentation — Foundation and Vocabulary.” n.d. International Organization for Standardization. Accessed December 11, 2023. https://www.iso.org/obp/ui/#iso:std:iso:5127:ed-2:v1:en:sec:3.1.6.14.
John, Nicholas A. 2013. “The Construction of the Multilingual Internet: Unicode, Hebrew, and Globalization.” *Journal of Computer-Mediated Communication* 18 (3): 321–38.

https://doi.org/10.1111/jcc4.12015.
Khanna, Saurabh, Jon Ball, Juan Pablo Alperin, and John Willinsky. 2022. “Recalibrating the Scope of Scholarly Publishing: A Modest Step in a Vast Decolonization Process.” *Quantitative Science Studies*, December, 1–43. https://doi.org/10.1162/qss_a_00228.
Kiaer, Jieun, Alessandro Bianchi, Giulia Falato, Pia Jolliffe, Kazue Mino, and Kyungmin Yu. 2022. *Missionary Translators: Translation of Christian Texts in East Asia*. Routledge Studies in East Asian Translation. London; New York: Routledge Taylor & Francis Group.
Klyshinsky, Edward, Vadim Maximov, and Sergey Yolkeen. 2008. “Cross-Language Transcription of Proper Names.” *Language Forum* 34 (1): 137–53. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=02539071&v=2.1&it=r&id=GALE%7CA258726534&sid=googleScholar&linkaccess=abs.
Konno, Ko, Munemitsu Akasaka, Chieko Koshida, Naoki Katayama, Noriyuki Osada, Rebecca Spake, and Tatsuya Amano. 2020. “Ignoring Non-English-Language Studies May Bias Ecological Meta-Analyses.” *Ecology and Evolution* 10 (13): 6373–84. https://doi.org/10.1002/ece3.6368.
Korpela, Jukka K. 2006. *Unicode Explained*. O’Reilly Media, Inc.
Kraker, Peter, Maxi Schramm, and Christopher Kittel. 2021. “Discoverability in (a) Crisis.” *ABI-Technik* 41 (1): 3–12. https://doi.org/10.1515/abitech-2021-0003.
Kris, Olds, and Susan L. Robertson. 2016. “Rankings as Global (Monetising) Scopic Systems.” In *Global Rankings and the Geopolitics of Higher Education*. Routledge.
Kulczycki, Emanuel, ed. 2023. “Playing the Evaluation Game.” In *The Evaluation Game: How Publication Metrics Shape Scholarly Communication*, 157–81. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009351218.007.
Kulczycki, Emanuel, Raf Guns, Janne Pölönen, Tim C. E. Engels, Ewa A. Rozkosz, Alesia A. Zuccala, Kasper Bruun, et al. 2020. “Multilingual Publishing in the Social Sciences and Humanities: A Seven-country European Study.” *Journal of the Association for Information Science and Technology* 71 (11): 1371–85. https://doi.org/10.1002/asi.24336.
Kupfer, Meital, and Jason Muyumba. 2022. “Language & Coloniality: Non-Dominant Languages in the Digital Landscape.” Pollicy. 2022. https://pollicy.org/resource/language-coloniality-non-dominant-languages-in-the-digital-landscape/.
Kwet, Michael. 2019. “Digital Colonialism: US Empire and the New Imperialism in the Global South.” *Race & Class* 60 (4): 3–26. https://doi.org/10.1177/0306396818823172.
Laakso, Mikael, and Janne Pölönen. 2023. “Mapping the Global Landscape of Journals.” Presented at the Helsinki Initiative Webinar on Multilingualism in Scholarly Communication, May 31. https://www.helsinki-initiative.org/en/events/helsinki-initiative-webinar-multilingualism-scholarly-communication.
Lacey, Eve. 2018. “Aliens in the Library: The Classification of Migration.” *KNOWLEDGE ORGANIZATION* 45 (5): 358–79. https://doi.org/10.5771/0943-7444-2018-5-358.
Leonard, Wesley Y. 2012. “Reframing Language Reclamation Programmes for Everybody’s Empowerment.” *Gender and Language* 6 (2): 339–67. https://doi.org/10.1558/genl.v6i2.339.
———. 2017. “Producing Language Reclamation by Decolonising ‘Language.’” *Language Documentation and Description*, December, Vol. 14 (2017). https://doi.org/10.25894/LDD146.
Lillis, Theresa M. 2010. *Academic Writing in a Global Context: The Politics and Practices of Publishing in English*. Routledge.
Liu, Esther Ruth. 2018. “The Role of Missionary Translation in African Colonial Politics.” In *The Routledge Handbook of Translation and Politics*, edited by Fruela Fernández and Jonathan Evans, 1st ed., 494–509. Routledge. https://doi.org/10.4324/9781315621289-33.
Luzón, María José. 2019. “‘Meet Our Group!’:” *International Journal of English Studies* 19 (2): 37–59. https://doi.org/10.6018/ijes.382561.

Makoni, Sinfree, Anna Kaiper-Marquez, and Lorato Mokwena, eds. 2022. *The Routledge Handbook of Language and the Global South/s*. 1st ed. London: Routledge. https://doi.org/10.4324/9781003007074.

Mamdani, Mahmood. 2019. “Decolonising Universities.” In *Decolonisation in Universities: The Politics of Knowledge*, edited by Jonathan Jansen, 15–28. Wits University Press. https://www.cambridge.org/core/books/decolonisation-in-universities/decolonising-universities/96E74F8E717D1B571920005B99F6EADB.

Matusiak, Krystyna K., Ling Meng, Ewa Barczyk, and Chia-Jung Shih. 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): 136–51. https://doi.org/10.1108/EL-08-2013-0141.

McKinney, Carolyn, Pinky Makoe, and Virginia Zavala, eds. 2023. *The Routledge Handbook of Multilingualism*. 2nd ed. London: Routledge. https://doi.org/10.4324/9781003214908.

Mikami, Yoshiki, and Kodama Shigeaki. 2012. “Measuring Linguistic Diversity On The WEB.” In *NET.LANG: Towards the Multilingual Cyberspace*, 118–39. https://digital.library.unt.edu/ark:/67531/metadc1743079/m2/1/high_res_d/netlang_EN_pdfedition.pdf.

Montgomery, Catherine, and Sheila Trahar. 2023. “Learning to Unlearn: Exploring the Relationship between Internationalisation and Decolonial Agendas in Higher Education.” *Higher Education Research & Development* 42 (5): 1057–70. https://doi.org/10.1080/07294360.2023.2194054.

Monyela, Madireng. 2021. “Call Us by Our Names: The Need to Establish Authority Control Standards for Non-Roman Names.” *Library Philosophy and Practice (e-Journal)*, June. https://digitalcommons.unl.edu/libphilprac/5516.

Morales, Esteban, Erin C. McKiernan, Meredith T. Niles, Lesley Schimanski, and Juan Pablo Alperin. 2021. “How Faculty Define Quality, Prestige, and Impact of Academic Journals.” *PLOS ONE* 16 (10): e0257340. https://doi.org/10.1371/journal.pone.0257340.

Morrisson, Heather. 2021. “Dysfunction in Knowledge Creation and Moving Beyond.” In *Global University Rankings and the Politics of Knowledge*, edited by Michelle Stack, 109–32. Toronto: University of Toronto Press. https://doi.library.ubc.ca/10.14288/1.0398205.

Naumova, Karina. 2014. “Legal Aspects of Transcription of Personal Names in the Latvian Language.” *RGSL Research Papers* 11. https://www.rgsl.edu.lv/uploads/research-papers-list/12/naumova-final.pdf.

Ndhlovu, Finex, and Leketi Makalela. 2021. *Decolonising Multilingualism in Africa: Recentering Silenced Voices from the Global South*. Multilingual Matters. https://doi.org/10.21832/9781788923361.

Nolan, Jason. 2006. “The Influence of ASCII on the Construction of Internet-Based Knowledge.” In *The International Handbook of Virtual Learning Environments*, edited by Joel Weiss, Jason Nolan, Jeremy Hunsinger, and Peter Trifonas, 207–20. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-3803-7_7.

O’Grady, William. 2018. “Assessing Language Revitalization: Methods and Priorities.” *Annual Review of Linguistics* 4 (1): 317–36. https://doi.org/10.1146/annurev-linguistics-011817-045423.

Olko, Justyna, and Julia Sallabank, eds. 2021. *Revitalizing Endangered Languages: A Practical Guide*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108641142.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. “OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts.” arXiv. https://doi.org/10.48550/ARXIV.2205.01833.

R’boul, Hamza. 2022a. “English and the Dissemination of Local Knowledges: A Problematic for South–South Dialogue.” In *The Routledge Handbook of Language and the Global South/s*, 144–57. Routledge.

R’boul, Hamza. 2022b. “The Spread of English in Morocco: Examining University Students’

Language Ontologies." *English Today* 38 (2): 72–79. https://doi.org/10.1017/S0266078420000449.
Riva, Pat. 2022. "The Multilingual Challenge in Bibliographic Description and Access." *JLIS*, no. 1. https://doi.org/10.4403/jlis.it-12737.
Rovira, Cristòfol, Lluís Codina, and Carlos Lopezosa. 2021. "Language Bias in the Google Scholar Ranking Algorithm." *Future Internet* 13 (2): 31. https://doi.org/10.3390/fi13020031.
Said, Edward W. 1979. *Orientalism: Edward W. Said*. 1st edition. New York: Ballantine Books.
Santos, Boaventura de Sousa. 2011. "Épistémologies du Sud." Translated by Magali Watteaux. *Études rurales*, no. 187 (August): 21–50. https://doi.org/10.4000/etudesrurales.9351.
———. 2018. *The End of the Cognitive Empire: The Coming of Age of Epistemologies of the South*. Durham, NC: Duke University Press.
Schmidt, Nora. 2020. "The Privilege to Select. Global Research System, European Academic Library Collections, and Decolonisation." Lund, Sweden: Lund University, Faculties of Humanities and Theology, Lund Studies in Arts and Cultural Sciences. https://doi.org/10.5281/zenodo.4302687.
———. 2023. "A Decolonial Approach to Open-Access Repositories: How to Set Up a Subject Repository for Documents on International Cultural Relations." Application/pdf. ifa (Institut für Auslandsbeziehungen e.V.). https://doi.org/10.17901/561.
Science Europe and COARA. 2022. "The Agreement on Reforming Research Assessment." https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf.
Shamma, Tarek. 2018. "Translation and Colonialism." In *The Routledge Handbook of Translation and Culture*. Routledge. https://web.unica.it/unica/protected/417116/0/def/ref/MAT417111/.
Shearer, Kathleen, Leslie Chan, Iryna Kuchma, and Pierre Mounier. 2020. "Fostering Bibliodiversity in Scholarly Communications: A Call for Action," April. https://doi.org/10.5281/zenodo.3752923.
Shearer, Kathleen, Silvia Mirlene Nakano Koga, Eloy Rodrigues, Natalia Manola, Martine Pronk, and Vanessa Proudman. 2023. "Current State and Future Directions for Open Repositories in Europe." Zenodo. https://doi.org/10.5281/ZENODO.10255559.
Siversten, Gunnar. 2018. "Balanced Multilingualism in Science." *BiD: Textos Universitaris de Biblioteconomia i Documentació* 40. https://doi.org/10.1344/BiD2018.40.25.
Souphavanh, Anousak, and Theppitak Karoonboonyanan. 2005. *Free/Open Source Software: Localization*. New Delhi: United Nations Development Programme-Asia Pacific Development Information Programme.
St Clair, Ralf. 2021. "Marginalizing the Marginalized: How Rankings Fail the Global South." In *Global University Rankings and the Politics of Knowledge*, edited by Michelle Stack, 133–49. Toronto: University of Toronto Press. https://doi.library.ubc.ca/10.14288/1.0398205.
Stack, Michelle, and Mayumi Ishikawa, eds. 2021. "Between Local Distinction and Global Reputation: University Rankings and Changing Employment in Japan." In *Global University Rankings and the Politics of Knowledge*, 153–71. Toronto: University of Toronto Press. https://doi.library.ubc.ca/10.14288/1.0398205.
Steigerwald, Emma, Valeria Ramírez-Castañeda, Débora Y C Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. "Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future." *BioScience* 72 (10): 988–98. https://doi.org/10.1093/biosci/biac062.
Stiller, Juliane, Vivien Petras, Maria Gäde, and Antoine Isaac. 2014. "Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences." In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, edited by Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen, and Ewald Quak, 238–47. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-13695-0_23.

"Supported Scripts." n.d. Unicode. Accessed December 11, 2023. https://unicode.org/standard/supported.html.
Tennant, Jonathan P. 2020. "Web of Science and Scopus Are Not Global Databases of Knowledge." *European Science Editing* 46 (October): e51987. https://doi.org/10.3897/ese.2020.e51987.
Torres, Nelson Maldonado. 2017. "Fanon and Decolonial Thought." In *Encyclopedia of Educational Philosophy and Theory*, edited by Michael A. Peters, 799–803. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-287-588-4_506.
Turner, Irina. 2023. "Decolonisation through Digitalisation? African Languages at South African Universities." *Curriculum Perspectives* 43 (1): 73–82. https://doi.org/10.1007/s41297-023-00196-w.
UNECSO. 2015. "A Decade of Promoting Multilingualism in Cyberspace - UNESCO Digital Library." CI-2015/WS/5. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000232743.
UNESCO. 2021. "UNESCO Recommendation on Open Science." UNESCO. https://doi.org/10.54677/MNMH8546.
"Unicode." 2021. EPrints Documentation. March 26, 2021. https://wiki.eprints.org/w/Unicode.
"Unsupported Scripts." n.d. Accessed January 5, 2024. https://unicode.org/standard/unsupported.html.
"Usage Statistics of Character Encodings for Websites." n.d. W3Techs: Web Technology Surveys. Accessed January 5, 2024. https://w3techs.com/technologies/overview/character_encoding.
Uzuner, Sedef. 2008. "Multilingual Scholars' Participation in Core/Global Academic Communities: A Literature Review." *Journal of English for Academic Purposes* 7 (4): 250–63. https://doi.org/10.1016/j.jeap.2008.10.007.
Vaughan, Crystal. 2018. "The Language of Cataloguing: Deconstructing and Decolonizing Systems of Organization in Libraries." *Dalhousie Journal of Interdisciplinary Management* 14 (April). https://doi.org/10.5931/djim.v14i0.7853.
Vera-Baceta, Miguel-Angel, Michael Thelwall, and Kayvan Kousha. 2019. "Web of Science and Scopus Language Coverage." *Scientometrics* 121 (3): 1803–13. https://doi.org/10.1007/s11192-019-03264-z.
Williams, Quentin, Ana Deumert, and Tommaso M. Milani, eds. 2022. *Struggles for Multilingualism and Linguistic Citizenship*. Multilingual Matters 173. Bristol; Jackson: Multilingual Matters.
Wu, Anping, and Jiangping Chen. 2022. "Sustaining Multilinguality: Case Studies of Two Multilingual Digital Libraries." *The Electronic Library* 40 (6): 625–45. https://doi.org/10.1108/EL-03-2022-0061.
Zaugg, Isabelle A., Anushah Hossain, and Brendan Molloy. 2022. "Digitally-Disadvantaged Languages." *Internet Policy Review* 11 (2). https://policyreview.info/glossary/digitally-disadvantaged-languages.