Resource Repositories and linking resources: An exploratory study

Setaka, Mmasibidi
South African Centre for Digital Language Resources
ORCiD: https://orcid.org/0000-0002-2790-3344
mmasibid.setaka@nwu.ac.za

Trollip, Benito
South African Centre for Digital Language Resources
ORCiD: https://orcid.org/0000-0003-2969-2832
benito.trollip@nwu.ac.za

Abstract

In this article the existence, use and importance of repositories are explored. An introduction into language resources (LRs) is given as well as a discussion of two platforms for the distribution of language resources, namely, the repository of the South African Centre for Digital Language Resources (SADiLaR) and Lanfrica, a site that links resources. In this article, types of repositories, such as institutional and language resource repositories, will be distinguished and compared. Language preservation is proposed as an important aspect which can be strengthened by the presence and use of repositories. The view expressed in this article is that the availability of language resources and repositories are pivotal for the development, preservation and advancement of languages.

Having a host site that links available resources and a repository where resources could be uploaded is a positive attribute of the mentioned online platforms, however as it will be discussed, the fact that information is available online is not a guarantee that the resources are or will be used by researchers or other interested persons, especially if they are not aware of their existence.

The article is concluded with suggestions for future work, for example measuring the influence of inaccurate metadata of language resources on linguistic research.

Keywords: language resources, language preservation, repositories, under-resourced languages, data

1 Introduction

Language resources (hereafter LRs) exist for many languages, more for some and less for others (Krauwer 2003). They are a recent development that occurred in the 2000s (Xie & Matusiak 2016) and allow for the acquisition, preparation, collection, management, and customization of datasets of different types, for example lemmatizers, corpora, dictionaries, parsers, and language identifiers. According to the European Language Resources Association (ELRA), LRs are inclusive of spoken and written corpora, computational tools, lexicographic resources and terminology databases. Resources such as morphological analysers, part of speech taggers, lemmatisers and so forth are developed from the collected corpora (spoken and written). Once these resources are collected, it is important to store them in a location that will allow for ease of access (Broeder et al. 2006), manipulation, updating and downloading. LRs have proved to be essential tools for research and development (Krauwer 2003; Itai and Wintner 2007), hence infrastructures such as the South African Centre for Digital Language Resources (SADiLaR), Lanfrica and Common Language Resources and Technology Infrastructure (CLARIN) have been established and mandated to host LRs so that researchers can find them. Repositories from these infrastructures can be used for a range of purposes, such as records management, research, learning, e-science, publication, and preservation. They can take on a variety of formats, such as e-print repositories, learning object repositories, and institutional repositories (Denison 2007).

Institutional Repositories (hereafter IRs) and language resource repositories (hereafter LRRs) are increasingly having an effect on the storage and accessibility of digital research data and they play a significant role in today's world. As a result of the availability of scholarly resources in digital formats, and in response to Open Access laws and regulations, IRs are sprouting all over the world, according to Kitchen & Mutiis (2016). This notion can also be alluded to by the LRRs, as more and more scholars are realising the importance of having a storage facility that can be accessed online.

UNESCO's initiative to work on Indigenous languages and recognise their value is a great way to prove how valuable language is (Bangani & Moyo 2019). It is for this reason that UNESCO and the Confederation of Open Access Repositories issued a statement together in support of the development of repositories with the aim of offering freely available research outputs via repositories (Kitchen & Muttis 2016). With this said, repositories have been playing a big role in preserving Indigenous Knowledge and languages. Kitchen & Mutiis (2016) indicate that the repository landscape in Africa seems the strongest in East, Southern and North Africa [1].

The purpose of this article is to explore IRs and LRRs, to investigate to what extent do LRRs more specifically offer benefits or whether they could possibly hinder the development, study and preservation of under-resourced languages in South Africa. We will firstly give a general background on existing literature on repositories, mainly to highlight some differences between IRs and LRRs. Thereafter we will illustrate how SADiLaR's repository and Lanfrica's search engine are similar and different with respect to the availing of African language resources. For the purpose of this article we consider the repository hosted by SADiLaR and the platform for linking resources hosted by Lanfrica as practical examples. Both have a focus on resources for languages found in Africa and include different LRs for different languages. SADiLaR's repository (2022) and Lanfrica's search engine (2022) will therefore be used as practical examples in this exploratory study.

The establishment and availability of these two platforms prove that the internet has been paramount in instilling a general pattern towards coordinating linguistic resources and storing them online (Broeder et al., 2006). This has made the importance of digital repositories even greater because they have the capacity to host a large number of resources in different formats (Vrana 2011).

We will conclude with a summary and directions for future work.

2 Literature review

Repositories offer a way to store data or research for future uses and the ability to access it digitally opens more opportunities and benefits (Hamid 2016). Lyon (2016) lists the opportunities as follows: it increases the possibility of research having more impact and visibility; it favours the reproducibility of science, it saves costs when creating data and it promotes and contributes to increased credibility in the system for researchers and scholars. Though sharing of data is now common practice, it is still a challenge because there are researchers who are not willing to share their research data (Gómez *et al.* 2016). Information systems called repositories absorb, manage, archive, and make accessible digital content (Xie and Matusiak 2016). Such artefacts include software, documentation, maps, information systems, and discrete manufactured components and systems. (e.g., electronic circuits, aeroplanes, automobiles, industrial plants). There are different types of repositories that can be found and distinguishable according to properties that they have. In trying to answer the question of what language repositories are. The Native Languages Archives Repository Project (Maynor *et al.* n.d.) defines a language repository as a collection of linguistic content that has been organised and made accessible to the public. The idea should be to have a centrally organised storage space for language resources in digital repositories.

According to Saini (2018), various repositories have been developed in the last two decades, and medium and small-sized institutions have also started planning and implementing them for scholarly support. Luarte (2006) mentions repositories such as consortia repositories, learning objects and discipline-based repositories (similar to subject based repositories). Subject or research repositories, also known as discipline repositories, focus on a certain field while based around a specific discipline (Armbruster and Romary 2010, Luarte 2006). The national repository system is intended to collect scholarly output more broadly, not simply to preserve a record of a certain field, and is much like an IR in that it caters for a broad spectrum of topics while also promoting teaching and learning in higher education (Armbruster & Romary 2010). These repositories function best in different settings, for example, the IRs which mostly function in academic settings and students, academics and librarians are responsible for them. IRs are differentiated from other repositories based on the types of services they offer which is mostly related to academic articles and electronic theses (Clobridge 2010). Additionally, IRs have the potential to improve access to and exchange of research-based data produced in Africa (Dlamini & Snyman 2017).

Foster (2008) mentions that it wasn't until late 2004 that IRs really gained momentum in the library world and only within the last few years have universities and colleges begun building IRs, either with open source software (and some with commercial support) like DSpace [2], Fedora [3], Greenstone [4], arXiv [5], or hosted services like bepress's Digital Commons [6].

Gibbons (2004a) argues that despite the many discussions and research that has taken place, it is quite difficult to provide a concrete and precise definition of an IR. This is the case because what is defined as a repository differs considerably, not only in Africa, but across the global scholarly landscape. This idea has therefore allowed for many definitions of IRs. Succinctly put, a repository is a location which is used to store data resources of different types and can be institutional or private.

With so many repository types available, it is important to raise awareness of LRRs as they serve communities at large and scholars within a certain discipline. They are of service not just to the organisation that hosts them; they add to the preservation of languages and cultural riches (Windhouwer et al. 2016). The growing trend of developing LRRs have helped in the collection and preservation of LRs and brought increasing efficiency and ease for data collection (Henke and Berez-Kroeker 2016). This is substantiated by the developments of SADiLaR's LRR and Lanfrica's linking website.

Low resource languages, African languages in particular, benefit from such resources as they often lack digital representation and scholarship, a point emphasised by Masakhane [7] - African languages are scarcely represented in technology. Once machine readable datasets are built, it is beneficial for them to be hosted in a LRR so that they can be discoverable and reused.

As already stated, we would like to focus on repositories geared towards LRs. The Linguistic Data Consortium [8] defines language resources as essential tools utilised by persons involved in language-related education, research, and technology development. The resulting materials that get developed are inclusive of data collections, corpora, software, research papers, and specifications which enable for the betterment and development of languages. Hence repositories began to host a variety of content (Xie & Matusiak 2016). ELRA [9] notes that these resources refer to machine-readable language datasets for building, improving, or evaluating natural language and speech algorithms or systems, or as core resources for the software localisation and language services industries, language studies, electronic publishing, international transactions, subject-area specialists, and end users.

Digital repositories remove the burden of authors to preserve and maintain their work themselves. While preserving and maintaining one's work can be a good thing, not depositing it in a repository often makes it undiscoverable for other scholars. A researcher preserving a conference paper on a computer's hard drive, for example, is generally responsible for its care and preservation (Gibsons 2004b).

2.1 FAIR and CARE principles of data

In a discussion on repositories one has to consider the FAIR and CARE principles. The FAIR principles (Findable, Accessible, Interoperable and Reusable) are a set of principles that emphasises the need for effective data stewardship. They serve as a guide to how data in all its forms should be findable, be accessible without restrictions, be able to be used in different formats and be reusable in order to allow other researchers to reproduce what has already been done instead of reinventing the wheel. (Wilkinson *et al.* 2016, Boeckhout 2018). While the FAIR principles are focused on data, there exists another set of principles that is more centred on people and purposes of the data as a result of the usage of data and the people whom it has been acquired from, especially the indigenous people (Carroll *et al.* 2020). These are the CARE Principles for Indigenous Data Governance (Collective benefit, Authority to Control, Responsibility, and Ethics) (Carroll *et al.* 2020; Gupta *et al.* 2020). These two sets of principles govern how data should be collected and distributed the question of ethics to the extent that Carroll *et al.* (2020) continues to argue that Indigenous Peoples must be represented and included in data processes that uphold ethical standards, as they will be the ones to weigh the advantages, drawbacks, and potential applications of data in light of local morals and values.

3 Elaboration on IRs, LRRs and a site that links resources to each other

When considering the digital preservation of languages or resources, online repositories are naturally an option (Masenya & Ngulube 2021). In Pinfield *et al.* (2014) the growth of more specifically open access repositories between 2005-2012 is discussed. In their paper the development of repositories in Africa is described as being "comparatively low" when compared to that in certain parts of Asia, Eastern Europe and South America (Pinfield *et al.* 2014:2415). While examining open access repositories in the BRICS countries that are findable on re3data.org, Misgar *et al.* (2022) concluded that South Africa has the least amount of open access repositories. More examples of studies about repositories include Pampel *et al.* (2013), who also discussed the visibility of repositories with reference to re3data, and Adam & Kaur (2021) that evaluate the functionality of IRs in Africa. The topic of IRs in a South African context, whether open access or not, therefore warrants more attention.

3.1 IRs in South Africa

Bangani & Moyo (2019) more specifically investigate the representation of African languages in IRs at public universities in South Africa. They provide a table summarising the content they found on the IRs with the University of Venda shown as hosting the IR with the highest percentage of content in an African language (5.75%) and the group of universities identified as Others (University of the Witwatersrand, University of the Free State, Walter Sisulu University, Sefako Makgatho Health Sciences University, Central University of Technology, Tshwane University of Technology, Vaal University of Technology, and Durban University of Technology) with no African language content in their repositories. The information in their table (Table 2 in their paper) is repeated here as Table 1 for reference (the only change is that the word *documents* in the heading of the second column has been shortened to *docs* to fit).

IRs, from a South African perspective at least, are therefore characterised by their association with or hosting by a university. Not only are they expressly linked to specific universities, but they normally contain outputs that are prototypical of universities, namely dissertations and theses (Banagani & Moyo 2019). Even though these resources are extremely important, it is crucial to consider where the data that theses and dissertations are based on, goes. The data in some studies require more stringent ethical safeguards, but that does not necessarily mean that metadata should not be made available for other researchers that could be interested in the study.

Table 1: Extent of African language representation in public university IRs in South Africa from Bangani &Moyo (2019)

Universities	Total no of docs	Total no of African language docs	% of African language docs
Cape Peninsula	3 606	3	0.08
Cape Town	27 255	7	0.03
Fort Hare	713	1	0.14
Johannesburg	29 562	2	0.01
KwaZulu-Natal	14 499	30	0.21
Limpopo	2 337	90	3.85
Nelson Mandela	4 961	22	0.44
North-West	27 201	7	0.03
Pretoria	54 508	58	0.11
Rhodes	10 448	4	0.04
Stellenbosch	54 497	39	0.07
Unisa	19 357	71	0.37
Venda	1 165	67	5.75
Western Cape	9 121	3	0.03
Zululand	1 624	68	4.14
Others	39 849	0	0
Total	300 961	472	0.14

For purposes of this article we will not consider IRs further, but shift the focus to LRRs. IRs that exclusively contain research outputs in the form of academic articles, theses or dissertations, like the IRs at South African universities and those discussed in Bangani & Moyo (2019), therefore fall outside the scope of the rest of this discussion.

3.2 A repository and a website that links sources

A prime example of a LRR, therefore not aimed at making theses and dissertations available, is SADiLaR's repository [10]. This repository contains a range of datasets and applications, downloadable or at the very least findable via relevant metadata or contact information of the responsible people or organisations. The focus in this repository is on South African languages and tools and resources developed for them. As of 29 August 2022 the repository contains a total of 406 assets.

Directly linking with the type of repository SADiLaR hosts, is Lanfrica - a linking site that is characterised as a search engine for African language resources. Importantly though, Lanfrica's inventory includes links to a broad range of research outputs, that subsumes the type of resources in IRs, sources in popular media, as well as resources found in LRRs. Resources are therefore not available for download on Lanfrica's website, but links to those resources are given. Lanfrica's scope is much broader than that of SADiLaR in as far as all African languages and their resources are relevant.

3.3 Benefits and challenges of LRRs

The benefits and challenges of a repository like SADiLaR's and a website like Lanfrica's should be considered. If we start with benefits, it is apparent that merely having a repository for African LRs can provide benefits in terms of findability, accessibility, interoperability and reproducibility as has been discussed earlier in this article and for CLARIN in De Jong *et al.* (2018). The nature of a repository like that of SADiLaR, which welcomes the submission of different kinds of language resources, provides a useful platform for the language and research community. The availability of resources is further augmented by providing links to websites where resources can be found, as in the case of Lanfrica's website. The fact that these two types of online mechanisms fulfil different, but complementary roles, is a benefit for researchers or other parties interested in online language resources. This benefit mainly manifests in terms of the findability of resources, especially considering the FAIR and CARE principles already discussed.

In terms of challenges, one has to consider the possible lack of awareness of the resources that are available and where to find them. It is not a given that the awareness or availability of resources guarantee their use or further development. The availability of information on repositories, their safety, and value should be emphasised in all engagements with the people or institutions that develop language resources. In Shode (2022) feedback on a recent natural language processing workshop is given. Shode mentions the linking of all the papers and datasets from the workshop on Lanfrica's site, in essence creating awareness and illustrating the practical benefit of Lanfrica's site. One could consider not only undertaking awareness campaigns or publishing blogs, but also tutorials illustrating the usage possibilities of LRs being discoverable and usable. A question that arises, and one that should be addressed in future work, is how could the availability of language resources add to the scale or type of linguistic research that is being done?

Following the lack of awareness as a possible challenge, one should also consider incomplete or inaccurate metadata. Metadata is of importance in the discussion of LRRs as it could lead to researchers not being able to find resources, even if they are aware of them. The most common characteristics of metadata that Park (2009) identifies are completeness, accuracy and consistency. Completeness pertains to the extent to which the metadata fields that are relevant to the specific type of resource meet the established requirements for the collection it belongs to. The accuracy of metadata refers to the correctness of the information given about the relevant resource, while consistency links with completeness in that there should be a consistent manner in which different types of sources are to be treated in the same collection. In their empirical study Park & Tosaka (2010) take a more practical approach to the criteria for measuring metadata quality. The survey they conducted was completed mainly by persons who are specialists in cataloguing and metadata (Park & Tosaka, 2010:703). Around three quarters of the respondents indicated accuracy as a measurement of data quality, while only about a quarter indicated currency (i.e. whether information contained in the metadata is current and latest). Park & Tosaka (2010:707, 711) comment on the low ranking of currency, stating that currency is important for making sources discoverable and there is a possible lack of systems or capacity for the continued maintenance of metadata leading to the low ranking.

One can therefore consider valid and current contact information for queries about sources as a priority where the metadata of assets in online repositories are concerned. The research community could experience inaccurate contact information as a barrier for engagement; it can lead to the data being overall inaccessible and the researcher becoming frustrated when queries go unanswered. A practical example of this is the contact person's email address given for the *SAE Pronunciation Dictionary* in SADiLaR's repository [11]. The record for this LR does not include a downloadable file and the email address given is not in use anymore, meaning interested researchers would be unable to easily inquire about receiving access to it.

A repository like the one offered by SADiLaR's is safeguarded through logins when parties want to submit resources. It is not only a barrier though, seeing as the approval of submissions also functions as a first step towards better quality assurance. Therefore adding an approval step to submitting a resource to SADiLaR's repository serves as a safeguard toward ensuring relevant sources are uploaded, rather than providing a platform where anything can be made available.

A challenge pertains to buy-in or participation by the research community. In this regard Woods & Pinfield (2022) can be considered. They examine studies that focus on incentivising the sharing of research or research data. Even though they focus on the context of data in the natural sciences, their approach and study could be extended to the humanities. From their summarising table, in which they specify the studies they include, some incentives mentioned are data sharing policies or mandates from funders and/or journals, cultural change and open data badges. The different incentives have differing degrees of effectiveness that ultimately illustrate how challenging this aspect is. A similar study for the types of data sharing incentives in and for humanities, as well as more specific with reference to the South African context, will definitely help with determining what is keeping institutes or researchers from sharing their research data and should be considered as future work.

4 Conclusion

In this article we discussed different aspects regarding repositories, mainly differentiating between IRs and LRRs. The possibilities for and challenges regarding the representation of African languages, also broadly within the frame of the preservation of languages, have also been included during the discussion. The general lack of repositories other than IRs that focus on prototypical outputs has been highlighted. SADiLaR's repository has been offered as an option for the depositing of LRs, seeing that its focus differs from IRs and it is aimed at making datasets, corpora and/or applications available. The nature of Lanfrica's search engine to link different resources has also been discussed and shown to be an asset for the language resource landscape. Observations from this exploratory article show what should be elaborated on in the future include the possible influence of LRRs on linguistic research in general, the lack of awareness of how LRRs could help, as well as how the sharing of data could be incentivised in a South African context.

Notes

[1] https://www.internationalafricaninstitute.org/repositories/news.phtml

[2] https://dspace.lyrasis.org/

[3] https://getfedora.org/

[4] https://www.greenstone.org/

[5] https://arxiv.org/

[6] https://bepress.com/products/digital-commons/

[7] https://www.masakhane.io/

[8] https://www.ldc.upenn.edu/language-resources

[9] http://www.elra.info/en/about/what-language-resource/

[10] http://repo.sadilar.org

[11] https://repo.sadilar.org/handle/20.500.12185/239

Acknowledgements

We would like to thank the reviewers of an earlier version of this article for their helpful comments and suggestions. Any views expressed are ultimately those of the authors.

Declaration of interests

It should be noted that both authors of this article are digital humanities researchers at SADiLaR. Care has been taken to keep the discussion and arguments focused and substantiated and limit any possible bias.

References

Adam, UA & Kaur, K 2021, Institutional repositories in Africa: Regaining direction. Information Development. pp 166-178 https://doi.org/10.1177%2F02666669211015429.

Armbruster, C & Romary, L 2010, Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. *International Journal of Digital Library Systems* (IJDLS), vol. 1, no. 4, pp.61-73. DOI: 10.4018/jdls.2010100104

Bangani, S & Moyo, M 2019, African Language Material in Institutional Repositories: Visibility and Scholarly Impact. *Mousaion*, vol. 37, no. 4. pp 1-20https://doi.org/10.25159/2663-659X/7237.

Boeckhout, M, Zielhuis, GA, & Bredenoord, A L 2018, The FAIR guiding principles for data stewardship: fair enough?. *European journal of human genetics*, vol. 26, no. 7, pp.931-936.

Broeder, D, Van Veenendaal, R, Nathan, D, & Stromqvist, S 2006, A grid of language resource repositories. *2006 Second IEEE International*

Conference on e-Science and Grid Computing (e-Science'06), Amsterdam, The Netherlands. pp1-5

Carroll, SR, Garba, I, Figueroa-Rodríguez, OL, Holbrook, J, Lovett, R, Materechera, S, & Hudson, M 2020, The CARE principles for indigenous data governance. *Data Science Journal*, vol. 19, no. 1, pp.1–12. https://doi.org/10.5334/dsj- 2020- 043/.

Clobridge, A, 2010, *Building a Digital Repository Program with Limited Resources*, Chandos Publishing, Oxford. pp272 https://doi.org/10.1016/B978-1-84334-596-1.50001-8.

De Jong, F, Maegaard, B, De Smedt, K, Fišer, D & Van Uytvanck, D 2018, CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Phoenix Seagaia Conference Centre, Miyazaki. pp3259-3264

De Mutiis, A & Kitchen, S 2016, African digital research repositories: Survey report. Africa Bibliography, 2015, vii-xxv. doi:10.1017/S0266673116000027.

Denison, T 2007. 'Library and information systems: a work in progress', in Ferguson, S (ed.) *Libraries in the twenty-first century: charting new directions in information services*, Chandos Publishing, Australia, pp. 165-177.

Dlamini, NN, & Snyman, M 2017, Institutional repositories in Africa: obstacles and challenges. *Library Review*, vol. 66, no. 6/7, pp.535-548.

European Language Resources Association (ELRA) n.d., 'What is a Language Resource?', viewed 10 May 2022, <http://www.elra.info/en/about/what-language-resource/>.

Foster, C 2008, Institutional Repositories - Strategies for the Present and Future, *DLTS publications,* Paper 4. pp1-12<http://digitalcommons.wku.edu/dlts_fac_pub/4>.

Gibbons, S 2004a, Defining an institutional repository. Library Technology Reports. 40. No. 4. pp.6-10.

Gibbons, S 2004b, Benefits of an institutional repository. Library Technology Reports. 40. No. 4. pp.11-16.

Gómez, ND, Méndez, E, & Hernández-Pérez, T 2016, Social sciences and humanities research data and metadata: A perspective from thematic data repositories. *El profesional de la información*, vol. 25, no. 4, pp.545-555.

Gupta, N, Blair, S, & Nicholas, R 2020, What we see, what we don't see: data governance, archaeological spatial databases and the rights of indigenous peoples in an age of big data. *Journal of Field Archaeology*, vol. 45 (sup1), S39-S50.

Hamid, B 2016, A model repository description language-MRDL. *International Conference on Software Reuse*, Limassol, Cyprus. pp.1-18

Henke, R & Berez-Kroeker, AL 2016, A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation and Conservation*, vol. 10, pp. 411–457.

International African Institute 2019, African Digital Research Repositories, viewed 12 May 2022 <https://www.internationalafricaninstitute.org/repositories/news.phtml>.

Krauwer, S 2003, The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of SPECOM*, Moscow, Russia. pp.1-8

Lanfrica 2022, viewed 9 May 2022, < https:lanfrica.com>.

Linguistic Data Consortium. (n.d.), 'Language Resources', viewed 10 May 2022, <https://www.ldc.upenn.edu/language-resources>.

Luarte, AL, 2006 Digital repositories : issues and challenges. Discussion Paper. (Unpublished). pp.1-39 <https://vuir.vu.edu.au/792/2/Setting_up_a_Repository.pdf >

Lyon, L 2016, Transparency: the emerging third dimension of open science and open data". *Liber* *quarterly*, vol. 25, no. 4. pp. 153-171 http://dx.doi.org/10.18352/lq.10113

Masenya, TM & Ngulube, P 2021, Digital preservation systems and technologies in South African academic libraries. *South African Journal of Information Management*, vol. 23, no. 1: a1249. https://doi.org/10.4102/sajim.v23i1.1249

Maynor, H, Cooper, S, & Shown Harjo, S. (n.d.), Native Language Preservation. A Reference Guide for Establishing Archives and Repositories. American Indian Higher Education Consortium. pp.279 Retrieved from http://www.aihec.org/resources/documents/NativeLanguagePreservationReferenceGuide.pdf.

Misgar, SM, Bhat, A & Wani, ZA 2022, A study of Open Access research data repositories developed by BRICS countries, *Digital Library Perspectives*, vol. 38, no. 1, pp. 45-54. https://doi.org/10.1108/DLP-02-2020-0012

North-West University 2022, Boloka Institutional Repository, viewed 9 May 2022 <https://repository.nwu.ac.za/>.

Pampel, H, Vierkant P, Scholze, F, Bertelmann, R, Kindling, M, et al. 2013, Making Research Data Repositories Visible: The re3data.org Registry. *PLoS ONE*, vol. 8, no. 11: e78080. doi:10.1371/journal.pone.0078080.

Park, J-R 2009, Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly,* vol. 47, no. 3-4, pp. 213–228. DOI: 10.1080/01639370902737240

Park J-R & Tosaka, Y 2010, Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms, *Cataloging & Classification Quarterly*, vol. 48, no. 8, 696-715, DOI: 10.1080/01639374.2010.508711

Pinfield, S, Salter, J, Bath, PA, Hubbard, B, Millington, P, Anders, JH and Hussain, A 2014. Open‐access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. *Journal of the association for information science and technology*, vol. 65, no. 12, pp. 2404-2421.

SADiLaR 2022, 'SADiLaR Language Resource Repository', viewed 29 August 2022, <https://repo.sadilar.org>.

Saini, O. P. (2018). The emergence of institutional repositories: a conceptual understanding of key issues through review of literature. *Library Philosophy and Practice*. pp.1-19

Shode, I 2022, 'We Ain't Just Cooking Jollof Rice, We Are Building AfricaNLP', view on 31 August 2022, <https://lanfrica.com/blog/we-aint-just-cooking-jollof-rice-we-are-building-africanlp/>

UNESCO (n.d.) Indigenouse Language Decade. view 14 May 2022, <https://en.unesco.org/idil2022-2032>.

Vrana, R, 2011, Digital repositories and the future of preservation and use of scientific knowledge, *Informatologia*, vol. 44, no. 1, pp.55-62.

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, & Mons, B 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, vol. 3, no. 1, 1-9. https://doi.org/10.1038/sdata.2016.18

Windhouwer, M, Kemps-Snijders, M Trilsbeek, P, Moreira, A, van der Veen, B, Silva, G, and von Reihn, D 2016. FLAT: Constructing a CLARIN Compatible Home for Language Resources. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. pp.2478-2483

Woods, HB & Pinfield, S 2022, *Wellcome Open Research*, vol. 6, no. 355. pp.1-23 https://doi.org/10.12688/wellcomeopenres.17286.2

Xie, I & Matusiak, K 2016, *Discover digital libraries: Theory and practice*. Elsevier, Amsterdam. pp.1-365