Aspects of a Legal Framework for Language Resource Management

Aditi Sharma Grover[1], Annamart Nieman[2], Gerhard B. van Huyssteen[3], Justus C. Roux[3]
Human Language Technology Research Group, CSIR-Meraka Institute, Pretoria, South Africa[1],
Advocate, Member of the Johannesburg Bar, Sandton, South Africa[2],
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa[3]
E-mail: asharma1@csir.co.za, nieman@law.co.za, gerhard.vanhuyssteen@nwu.ac.za, justus.roux@nwu.ac.za

Abstract

The management of language resources requires several legal aspects to be taken into consideration. In this paper we discuss a number of these aspects which lead towards the formation of a legal framework for a language resources management agency. The legal framework entails examination of; the agency's stakeholders and the relationships that exist amongst them, the privacy and intellectual property rights that exist around the language resources offered by the agency, and the external (e.g. laws, acts, policies) and internal legal instruments (e.g. end user licence agreements) required for the agency's operation.

Keywords: Africa, legal, language resource management, resource-scarce

1. Introduction

During a workshop on legal aspects of electronic language resources at the 2010 Language Resource and Evaluation Conference in Malta, various speakers expressed the need for a better understanding of the legal frameworks, both generic and country specific, governing electronic language resources. This article aims to investigate some of the aspects of such legal frameworks; while we will generalize away from the country-specific aspects, our investigation specifically stems from the establishment of a resource management agency (RMA) by the South African governments' Department of Arts and Culture. This RMA will be responsible for the management (i.e. collection, curation, warehousing, and distribution) of resources of South African languages, similar to the activities of agencies like the Dutch TST-Centrale, the European ELRA/ELDA, the USA's LDC, etc. (cf. Roux et.al 2011 and Roux 2011).

RMAs like these operate within a legal framework that formalizes the relationships with stakeholders that provide or use the language resources (LRs) that a RMA manages:

- **Stakeholders**: various entities that are directly and indirectly involved in the operations of a RMA, including content providers, service providers, partners, etc. (section 2);
- **Language resources**: the objects that serve as the core responsibility and offering of the RMA (section 3);
- **Legal framework**: legal instruments external to a RMA (e.g. laws, treaties, etc.) and legal instruments internal to a RMA (e.g. license agreements, contracts, service level agreements, etc) (section 4).

2. Stakeholders

One of the first steps in defining the legal framework is to *identify the priority relationships of a RMA, which need to be formalised by legal means;* this is done through a stakeholder analysis of a RMA. These stakeholders include:

Primary content providers: These are providers of corpora, lexica and technologies (i.e. language models, software, etc.) for management by a RMA. In South Africa, the majority of such resources are provided by agencies involved in projects commissioned by Government, although other institutions might also voluntarily contribute their resources on a need-to basis.

Secondary content providers: These are content providers that indirectly contribute language data usually and preferably via primary content providers (since the primary service providers are the ones commissioned by government to collect resources). The relationship between the RMA and this category of content providers is mostly regulated through data release agreements between them and the primary content providers. Secondary content providers could include, *inter alia*, commercial entities (small-medium enterprises, publishers and corporates), governmental entities, the World Wide Web (WWW) and various individuals, in both amateur and professional capacities.

Service providers: Any RMA could be serviced by a number of external service providers, offering data storage/hosting, legal advisory, evaluation and validation services, etc.

End-users: Although the end-users of a RMA could typically include the primary and secondary content providers, its client base should ideally be more wide-ranging and diverse, including commercial entities, international organizations, other RMAs, etc.

Networks: In order to make full use of Internet-based national and international expertise, best practice, re-usable resources and computational tools, as well as infrastructures, it is imperative that the RMA links up with existing networks and professional organisations in the field. This linkage could be through informal or formal agreements with strategic partners (such as other RMAs, distribution agencies and standardisation organisations), and/or by participating in national and international initiatives and networks.

In terms of managing these various stakeholder relationships it is pivotal to *conduct a relationship audit for any to-be-established RMA, in order to get a full overview of all existing relationships, whether legally formalised or no.* Part of such a relationship audit is to

secure all original supporting legal documentation that should be kept in a *comprehensive contract and relationships register* for, *inter alia*, relationship, contract and rights management purposes. Care must be taken to uncover all the potential third parties involved in each relationship.

3. Language Resources

A RMA should *identify the priority LRs (i.e. HLT objects; cf. Sharma Grover et al. 2011) that are protected and/or to be protected by legal means*. Careful consideration is required when using the definitions for HLT objects (e.g. "corpora", "lexica", and "databases") in a legal context. Domain-specific, technical definitions within legal documents must be as clear, concise and technology-neutral as possible and most importantly, must be used consistently (preferably so not only within its own context but also within the contexts of other national and international legal instruments).

An *updated and comprehensive IP Register* is vital to the operations of a RMA. For the purpose of this task, *prior LR audits* (e.g. Sharma Grover *et al.* (2011), Binnenpoorte *et al.* (2002), Maegaard *et al.* (2009)) *can prove to be very valuable* in expediting this process. It is critical that the *IP arrangements underscoring the development of the priority LRs are neatly ironed out as this will constitute the due diligence basis upon which further LR development* will take place. In particular, it is important to note that the Internet/ WWW is often used to source corpora in LR generation (e.g. data hounds and crowd sourcing). From a risk management perspective, the RMA must appreciate that *various projects that mined the WWW for content will require due diligence scrutiny*.

4. Legal Framework

The most important legal rights that come into play with respect to the provision of content to the RMA include the privacy rights (section 4.1) and the IP rights (section 4.2), not only of the content providers but also of third parties with respect to the content and the use thereof. For purposes of illustrating the application of various legal rights to the RMA context, we will focus on the use of end user licence agreements (EULAs) by a RMA (section 4.3).

4.1 Privacy Rights

With regard to the privacy concerns underscoring content to be used by a RMA, cognisance should specifically be taken of the host country's specific legislative impetuses.

A general right to privacy could potentially be associated with content exploited by the RMA and could give rise to, *inter alia*, infringement liability. The importance of due diligence and resulting legal risk management in respect of IP resources cannot be overstated.

It is furthermore important to note that the processing of personal (such as a person's name, age, language, etc) and sensitive information (such as a person's religion, philosophy of life, race, political persuasion, health, etc.) calls for heightened protection and is generally more jealously guarded in law. The risks inherent in the processing of personal information (as, for example, defined in section 1(2) of the draft South African Bill on the Protection of Personal Information ("the PoPI Bill") include that the data may not be accessed or disclosed without authorisation and may not be used for a purpose other than that for which they were collected (*cf., inter alia*, article 25 of the European Union's Data Protection Directive 95/46/EC on the Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of such Data; and the South African "Regulation of Interception of Communications Act" (RICA) that deals with aspects that are dealt with in the European Union's Directive on the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector 2002/58/EC).

4.2 IP Rights

With regard to IP rights, the CLARIN Work Package 7 (http://www.clarin.eu/wp7/a-short-outline-of-the-work-package-7) rightfully points out that if LRs were free of copyrights and other restrictions, their sharing and use would be much simpler. The reality is, however, that although language *per se* is not subject to IP protection, most of the LRs and associated technology is governed by various restrictions in their copying, their showing in public and their use for specific purposes. However, from a pragmatic perspective, it should be pointed out that aside from content subject to various IP restrictions, a substantial amount of content could be exploited by a RMA because it belongs to the public domain. These public domain texts can usually be published or copied, mostly subject only to acknowledgement of the source. Notwithstanding, caution should be exercised when tagging content as squarely sitting in the public domain.

In addition, relating specifically to Africa, it should be noted that the traditional and indigenous knowledge and traditional cultural expressions or folklore do not fit easily into existing IP systems. During 2004, the South African government adopted an indigenous knowledge systems ("IKS") policy, which is considered an example of the kind of *sui generis* IP measure African nations are encouraged to institute. The policy attempts to find a balance between respecting and protecting tradition on the one hand and enabling community economic development through commercial use on the other. Thus, African IKS must be duly considered when IP arrangements are devised for the South African RMA.

The contract and (digital) IP rights management enabling the utilisation of these resources poses stark challenges and constitutes the important rationale for *a comprehensive IP due diligence audit to be conducted by a RMA*. The secondary underlying principle, of course, is that such a due diligence audit will position a RMA to protect its own IP rights going forward, should it opt to do so. Open source IP is specifically focused on in the sections below due to its topicality (sections 4.2.1 and 4.2.2).

4.2.1 Open Source IP domain

In this section we provide a high-level *overview of the open source IP domain, specifically from the perspective of the RMA* vis-à-vis that of its end-users. The open access movement in scholarly communication, the free/libre open source software (FLOSS/FOSS) movement, and the open content approach to online sharing and collaboration among authors are preeminent in this regard, and are briefly considered below.

Open access initiatives revolve mostly around the practice of academics making their research outputs and writings available on the Internet either through open access online journals (such as First Monday), online institutional archives (such as Brewster Kahle's Internet Archive) or online repositories (such as the repositories of academic institutions and libraries).

Although the definitions of free software and open source software have much in common, they differ in rhetoric, which reflects their differences in philosophy. Despite these differences, however, from a pragmatic perspective, the Free Software Foundation (FSF) and the Open Source Initiative (OSI) typically agree on the classification of FOSS and non-FOSS licences in most instances. There are currently sixty-seven OSI-approved licences and the list is increasing. A useful comparison of the most popular OSI-approved licences and its compatibility can be accessed through: http://www.openfoundry.org/en/comparison-of-licenses?tmpl=component&print=1&page.

The open content movement encourages online adaptation of materials by users. The Wikipedia collaborative encyclopaedia and the Creative Commons ("CC") licensing system (www.creativecommons.org) are the best-known open content projects. The CC Public Licences are inspired by the FOSS development and advocate for openness of all kinds of digital content such as music, literary texts, art works and photographs. CC licences, however, do not apply to software *per se* (although the CC licences are also used for software, the GNU GPL is considered the most well-known, comprehensive and suitable to the software licensing context).

There are currently six main CC licences (11 licences from a previous CC version are still available) which take into account four conditions relating to attribution, non commercial use, derivative works and sharing. The attribution requirement has become default since the requirement of attribution has been widely adopted by users of CC licences. The CC flexible licensing system allows authors to adopt a "some rights reserved" approach to their works. When using a CC licence, the author or creator specifies which uses he or she will allow other to make of his or her work and attaches the appropriate CC licence to the work online (thus providing copyright clearance to certain uses upfront as a tag to the file on the Internet). The CC Developing Nations licence allows an author to specify freer terms of use in the developing world that in developed nations, thus allowing an author to participate first-hand in reforming global policy. South Africa is currently the only African country to have "ported" the CC licences into its national jurisdiction, with the launch of the CC SA licences in May 2005.

It is also important to note that Creative Commons make available a public domain mark ("PDM") for labelling works that are free of known copyright restrictions. The PDM is intended for use with old works that are free of copyright restrictions around the world, or works that have been affirmatively placed in the worldwide public domain prior to the expiration of copyright by the rights' holder. Should an author want to free her own work of copyright restrictions, the CC0 public domain dedication is available for use.

4.2.2 External Legal Instruments for Open Source IP Domain

Several external legal instruments (such as laws, treaties, conventions, etc.) exist that affect the open source IP domain. A RMA will have to decide which instruments are most important for its legal framework; here we list a few examples relevant to the South African context. (For a comprehensive list, see Roux *et al.*, 2010.)

Examples of International Legal Instruments

- The Berne Convention for the Protection of Literary and Artistic Works of 1886.
- The Agreement on Trade-Related Aspects of Intellectual Property Rights (the TRIPS Agreement) concluded on 15 April and entered into force on 1 January 2005.
- The various TRIPS Plus arrangements in Free Trade Agreements (FTAs) with certain countries.
- The World Intellectual Property Organisation ("WIPO") Treaty ("WTO") adopted at the WIPO Diplomatic Conference on Certain Copyright and Neighboring Rights Questions that entered into force on 6 March 2002.
- The WIPO Performance and Phonograms Treaty ("WPPT").
- The Internet Treaties (i.e. the WTO and the WPPT referred to together).
- The WIPO Digital Agenda adopted in 1999.
- The Copyright Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright Law in the Information Society ("the Copyright Directive").
- The Directive on Certain Legal Aspects of the Information Society Services, in Particular Electronic Commerce, in the Internal Market 2000/31/EC (OJ L178 of 17 July 2000).
- Directive 98/34/EC of the European Parliament and of the Council of 22 June 1998 Laying Down a Procedure for the Provision of Information in the Field of Technical Standards and Regulations.
- The EU Council Directive on the Legal Protection of Databases adopted on 11 March 1996.
- Country-specific legislation (such as the United States Digital Millennium Copyright Act of 1998; the German Informations- und Kommunikationsdiente-Gesetz of 1997; the Botswana Copyright and Neighboring Rights Act of 2000; the United States Patent Act 35 USC and the United Kingdom Patents Act of 1977).
- The European Patent Convention.
- The Paris Convention for the Protection of Industrial Property of 20 March 1883.
- The Centre for Strategic & International Studies ("CSIS") updated their Open Source Policy Survey in March 2010. The Survey takes a worldwide look at Governmental Open Source Policies and divides them into four categories, namely research, mandate, preference and advisory. In total the CSIS found 364 open source policy initiatives worldwide. The CSIS Report not only considers the individual initiatives of each country, but also categorises the countries into regional group initiatives. The CSIS Report can be accessed at http://csis.org/files/publication/100416_Open_Source_Policies.pdf.
- Open Source Software has been recognised by SADC in the "Resolution of the SADC Parliamentary Forum (SADC PF) Information and Communication

Technologies (ICT's) Conference on 'e-Parliament, Concepts, Policies and Reality'" in October 2009. The SADC Resolution can be accessed at http://www.parliaments.info/documents/eparliament-resolution.

Examples of South African Legal Instruments

- The Copyright Act 98 of 1976.
- The Electronic Communications and Transactions Act 25 of 2002 ("the ECT Act").
- The Intellectual Property Rights from Publicly Financed Research and Development (IPR-PRFD) Act 51 of 2008.
- Policy on Free and Open Source Software use for South African Government, Cinematography Act 1977.
- Intellectual Property Laws Amendment Act 38 of 1997.
- Intellectual Property Laws Rationalisation Act 107 of 1996.

4.3 Internal Legal Instruments: EULAs

As part of setting up a legal framework of a RMA one will have to formalise various internal legal instruments, such as end user license agreements (EULAs), terms of references (TORs), service level agreements (SLAs), etc. For purposes of this publication and to illustrate the application of various legal rights within the RMA setting, we focus on a few issues related to EULAs.

EULAs form the back-bone of a RMA's operations. The formulation of a EULA is based on a RMA's business model, goals and objectives. A proper due diligence audit on the current priority LRs would be also required as this will constitute the basis for the EULA negotiations. In addition, the prescribed liaisons with local/international regulatory/legal bodies (e.g. in South Africa the National Intellectual Property Management Office ("NIPMO")) would also be instructive to the drafting of a EULA.

A multitude of questions and/or concerns should be taken into account when a RMA formulates/selects its EULA model, some of which we highlight below.

Rights conferred on users

The EULA must reflect the goals, aims, business model and profit generating mechanisms of the project (if any). In so doing, the EULA should reflect the decisions of the RMA/ proprietor of the IP regarding the following:

- *Attribution required?* Does the RMA/ proprietor require any person (user) that copies, distributes, displays, or performs the IP to credit the author or the RMA/proprietor? If yes, what form must such attribution take (i.e. should the user include a hyperlink to the RMA/ proprietor's website)?
- *Commercial or non-commercial uses permitted?* Does the RMA/proprietor permit persons to copy, distribute, display, and perform the IP for commercial or non-commercial purposes? Does the author distinguish between commercial and non-commercial uses for the different categories or types of available IP (i.e. whereas Version 1 is to be used for commercial purposes, the use of Version 2 is subject to the payment of a licence fee)?
- *Are adaptations of the IP permitted?* Are users at liberty to alter, transform, or build upon the IP and create adaptations (or modifications)?
- *Is distribution or sharing of the IP permitted?* May users distribute copies of the IP? If yes, are there any restrictions on distribution (such as within certain organisations and/or communities only)?

Due diligence

This refers to conducting due diligence of the licences of contributing authors to the final IP. Taking into consideration that IP derived from FOSS development is normally a collective effort between many different authors, alternatively a collection of the IP of various authors, a critical step in deciding on the appropriate licence is to determine the existing terms and conditions regarding the use of the IP of contributing authors and/or collaborators.

- Do their licenses have terms that could conflict with the RMA's choice of license? Who will ultimately own the rights in the final IP?

Proprietary vs. open licence/Proprietary & open licences

- Is it possible for a single organisation to license different products in different ways?

Software patent infringement

Countries such as the United States allow software to be patented (which is not the position in South Africa currently). This however does create the risk that an aspect of the FOSS code could be patented by another company. A licence that is incompatible with such consideration may result in patent infringement.

Trademark protection

Trademarks identify and distinguish products and services from those of third parties and all EULAs must deal with the manner of use of the RMA/proprietor's trademark (for example, that the trademark may not be removed from the licenced product).

Warranties

All licences must address the issue of limitation of liability for losses or damages suffered by the user. This is of particular significance in FOSS licences where the author of the original work cannot be held liable for the adaptations and modifications of the IP or the contentions in respect of, for instance, fitness for purpose, made by a distributor of the IP.

5. Conclusion

In this contribution we highlighted some of the aspects that need to be kept in mind when formulating a legal framework in which a RMA could operate. We provided broad categories of aspects that should be considered, viz. stakeholders (i.e. clients), language resources (i.e. products), and legal instruments (i.e. legislation, contracts and licences). Of course, for each specific context (e.g. country/region, language, etc.), specifics of that context will have to be considered, and need to be formulated before establishing a RMA. We hope, however, that this publication will help to guide other institutions in thinking about the legal frameworks of their to-be-established RMAs.

6. Acknowledgements

The authors would like to acknowledge the indirect inputs of various members of the DAC's HLT Expert Panel, as well as the financial contribution of DAC towards this investigation.

7. References

Sharma Grover, A, Van Huyssteen, GB & Pretorius, MW. (2011). The South African Human Language Technology Audit. Language Resources and Evaluation. DOI: 10.1007/s10579-011-9151-2. ISSN: 1574-020X. 45(3).

Binnenpoorte, D., De Vriend, F., Sturm, J., Daelemans, W., Strik, H., & Cucchiarini, C. (2002). A field survey for establishing priorities in the development of HLT resources for Dutch, In Proc. LREC 2002, Las Palmas, Spain, 1862–1866.

Maegaard, B., Krauwer, S., & Choukri, K. (2009). BLaRK for Arabic. MEDAR—Mediterranean Arabic Language and Speech Technology. http://www.medar.info/MEDAR_BLARK_I.pdf. Accessed June 2009.

Roux,JC, Van Huyssteen, GB, Gumede, T & Mojapelo, ML.(2010). Blueprint for NCHLT Resource Management Agency.

Roux,JC, Van Huyssteen, GB, Gumede, T & Mojapelo, ML.(2011) The South African National HLT Resource Management Agency. Poster. FlareNet Forum 2011, Venice, http://www.flarenet.eu.

Roux, JC. (2011) Developing language resources in an African context: The South African case. Position paper and presentation – Session 6. FlareNet Forum 2011, Venice, http://www.flarenet.eu.