Practical and Technical Aspects for Using the International Standard Language Resource Number

Khalid Choukri, Victoria Arranz, Olivier Hamon, Jungyeul Park

ELDA - Evaluations and Language resources Distribution Agency
55-57, rue Brillat-Savarin 75013 Paris, France
{choukri, arranz, hamon}@elda.org

Abstract

In this document, we propose a new unique and universal identification schema for Language Resources to provide Language Resources with unique names using a standardized nomenclature. This will also ensure Language Resources to be identified, and consequently to be recognized with proper references in activities within Human Language Technologies as well as in documents and scientific papers.

Keywords: International Standard Language Resource Number, ISLRN, Language Resources

1. Introduction

Every object in the world requires a kind of identification to be correctly recognized. Traditional printed materials like books, for example, have generally used the International Standard Book Number (ISBN), the Library of Congress Control Number (LCCN), and now in the digital world, the Digital Object Identifier (DOI) and several other identifiers as a unique identification scheme. Book identifiers allow us to easily identify books in a unique way. Other domains make use of several other identifier schemes. For instance, it is not hard to come into contact with an International/European Article Number (EAN), which is a universal bar-coding system for everyday products. Each of these schemes seem to have been the output of some specific need or circumstance within a domain.

After having reviewed extensively existing identification schemas[1] we conclude for the need to establish a specific identifier for LRs. Hence, we made a proposal for introducing a new identifier scheme for language resources (LRs), namely, the International Standard Language Resources Number (ISLRN). It is meant to provide LRs with unique identifiers using a standardized nomenclature. This will ensure that LRs are correctly identified, and consequently, recognized with proper references for their usage in applications in R&D projects, products evaluation and benchmark as well as in documents and scientific papers. Moreover, it is also a major step in the networked and shared world that Human Language Technologies (HLT) has become: unique resources must be identified as they are and meta-catalogues need a common identification format to manage data correctly. Therefore, LRs should carry identical identification schemes independently of their representations, whatever their types and wherever their physical locations (on hard drivers, Internet or Intranet) may be.

2. International Standard Language Resource Number (ISLRN)

For many different reasons, a LR may be duplicated (on different catalogues/databases), renamed, modified, moved, or deleted. Thus, a permanent and unique identifier associated to a LR will always permit to retrieve it. Furthermore, having the ISLRN requires also the building of the ISLRN centers that would manage their attribution, storage, and consistency.

The unique identifier does not intend to replace local and specific identifiers. For instance a resource that is distributed by several data centers will still have the local data-centre identifier but will have a unique ISLRN. For instance a resource distributed by ELRA and LDC will be identified in respective catalogues as e.g. ELRA-S0064 as well as LDC2008L01. It is our goal to introduce the universal and unique identifier so such resources (with different local references) can be seen by users as identical.

3. Principle of ISLRN

Most of the rationales, technical and scientific arguments were developed in (Choukri et al., 2011) and will not be duplicated herein. While reviewing existing Identifiers, we realized that in many areas, some semantics are included in the identifier e.g. publishers for books, year of publication, category of language resources in data centre catalogues. In such fields, clear ontology seems to be available and widely adopted. In our field, this seems to be very controversial given the lack of a unique ontology as shown by the information compiled within the LREC MAP (`http://www.resourcebook.eu`[2]). At the various community meetings, the option to associate content information with the identifier has been discarded. It was agreed that LR identifier should delegate semantics of its content to metadata which can easily and richly describe it. Therefore, we decide to use 10-digit random numbers as the new LR identifier followed by 2-digit for version information and 1-digit for a checksum number. The checksum number is encrypted from the preceding numeric identifier and version information. Our proposal is summarized in Figure 1.

[1] (Choukri et al., 2011).

[2] (Calzolari et al., 2010).

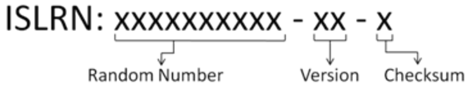

Figure 1: Schema proposal for the ISLRN

4. Attribution of ISLRN

The challenge behind the attribution of ISLRN is to ensure that key players within the field, highly representative and well respected are the guarantees of its sustainability as well as its free availability, and its ethical modus operandi. It is therefore essential that the principle and the operational aspects are endorsed by major players and data centres, acting as an informal “umbrella” organisation[3] (steering committee). It is also crucial that the ISLRN attribution is trusted to a small group of organizations involved in all LRs distribution and sharing issues. Such small executive committee (ISLRN Attribution body[4]) will have to set up an ISLRN server and to moderate the process, ensuring that all requests for ISLRN introduced by the community members are reviewed and given a quick reply.

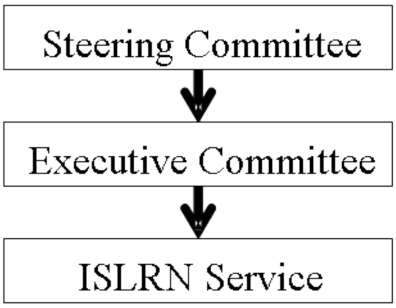

Figure 2: Proposal for the ISLRN management structure

We assume that most of the requests will be deemed eligible for an ISLRN. Cases that will have to be rejected are requests for attributions of ISLRN to objects that are indisputably out of the HLT scope or LR that have already been assigned an ISLRN.

In order to introduce a valid request for an ISLRN, the requester has to allow for a prerequisite checking before assigning the ISLRN. To carry such checks, LR Right holders or creators should provide minimum information that describes their LRs. Such “metadata[5]” will be requested during the first phase of ISLRN attribution. By submitting the LR and its description, the request will be checked by a moderator from the executive body. If the validation process by the moderator is affirmative, the ISLRN will be assigned; if the validation process fails, the ISLRN will not be assigned. The requester can appeal on such decision.

The synopsis of the ISLRN process and server of the first phase is depicted in Figure 3.

Phase 1 will be mainly focusing on the setting up of a stand-alone of a networked ISLRN server to which requests should be put. The requester will be responsible for filling all metadata elements required to clearly identify the LR and check that it has not been assigned an ISLRN yet.

A second phase is planned and will liaise with the major LR repositories that may have already catalogued and described the LR. The requester could indicate in which catalogue such LR is described and hence its metadata could be automatically harvested as described in Figure 4.

ISLRN should be assigned for free worldwide. The organizations endorsing the ISLRN attribution commit and will ensure that it is assigned for free without any entry fee or annual subscription. Since the ISLRN will not be a legal deposit, the ISLRN is not an obligation, but rather an essential and best practice.

5. ISLRN Infrastructure

As indicated above, the ISLRN management infrastructure will be developped over two phases. Phase 1 will allow to set up a standalone server (or network of mirrored servers) that manage all functionalities of the assignements of ISLRN and the corresponding backoffice. Phase 2 will ensure that the ISLRN server is connected and synchronised with the major data centers’ catalogues to make it easy for the ISLRN requesters to refer to resources that are already included in existing catalogues instead of filling the ISLRN forms in detail.

We provide here a quick descriptionn of some functionalities that the ISLRN infrastructure must include in phase 1 and phase 2.

5.1. Assignment procedure

The basic function of the server is to receive requets for data providers or rightholders (owners, distributors, compilers, etc.) through a user-friendly interface. Such interface should allow to receive requests for one LR or for a bunch of resources. The main items to be handled at that stage are:

- Submit LR metadata for an ISLRN assignment including details of the requester.
- Share the information collected between the ”assignment” team (at least one member from each party within the executive committee).
- The moderator reviews the Metadata collected and the associated request, and decides on the submission in accepting or rejecting the query.
- Notify the requester either by providing an ISLRN or a justification of the refusal.
- Add the LR and its metadata in the ISLRN database if the LR is assigned an ISLRN.

During that phase, all users could browse through the ISLRN server and look for an existing metadata corresponding to a LR in the ISLRN database. This may help

[3] Major players expected to be involved ACL, AFNLP, ALAGIN, ELRA, IAMT, ISCA, LDC, Oriental-Cocosda, etc.

[4] At this stage the executive body is expected to consist of at least ALAGIN, ELRA, LDC.

[5] Regarding to the list of metadata categories and components to describe LRs, OLAC metadata (Open Language Archives Community) is the easy choice at this stage.

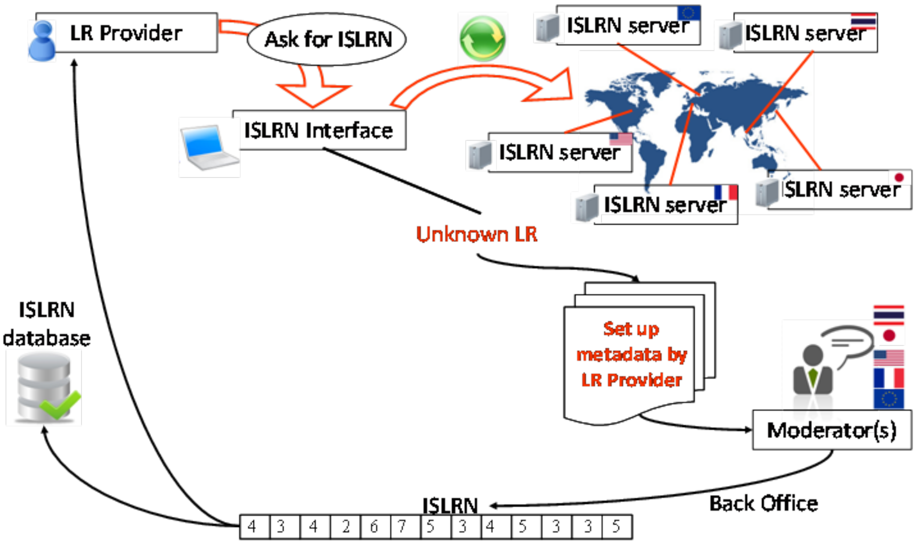

Figure 3: ISLRN attribution approach during phase 1

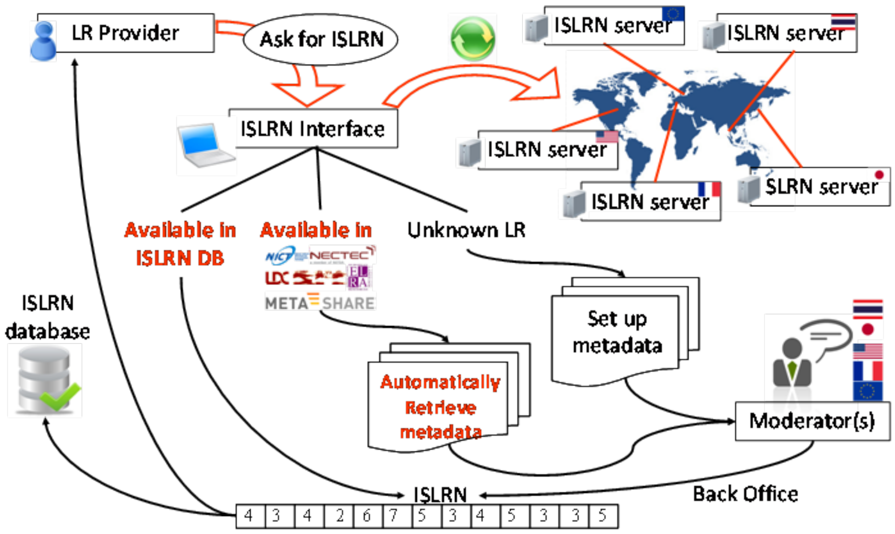

Figure 4: ISLRN attribution approach during phase 2

them to identify resources that have already been assigned an ISLRN and avoid duplicate requests.
During phase 2, the main items to be handled are those of phase 1 with additionnal ones that allow to interact with existing catalogues that comply with the ISLRN harvesting schema so that one would have:

- Submit a request to have an ISLRN assigned to a LR.
- Collect information on the requester.
- Collect information on which data center such LR is catalogued and its local reference.
- Harvest the metadata from that catalogue and ask the requester to confirm it. If the information is accurate then the ISLRN committee follows tasks of phase 1.
- Share the information collected between the ”assignment” team (at least one member from each party within the executive committee)
- The moderator reviews the Metadata collected and the associated request, and decides on the submission by accepting or rejecting the query.
- Notify the requester either by providing an ISLRN or a justification of the refusal.
- Add the LR and its metadata in the ISLRN database if the LR is assigned an ISLRN.
- Notify the corresponding data center that the LR has been assigned an ISLRN.

5.2. Management of the ISLRNs or the Back-Office

The role of the back office is to interact with the ISLRN requesters but also to conduct all reviews and necessary investigations before the attribution of the identifier. The main challenge is to ensure that a LR gets one ISLRN regardless of 1) who introduces the request and 2) the number of requests. The main tasks are:

- Review the metadata that has been submitted in support of the request.
- Interact with the requesters to get more info if needed
- Check that the LR was not assigned an ISLRN and that it is not under investigation to get one (no other request has been filled before).
- Intercat with the other moderator if needed to ensure that such LR could get an ISLRN.
- Assign through the associated server procedure a unique identifier for the LR. The server encrypts the number to avoid an ISLRN creation by entities not part of the ISLRN network.
- Notify the requester of the assigned ISLRN and update the ISLRN public database (synchronise all networked servers).

During phase 2 of the process, the ISLRN server will interact with the data centers to harvest metadata on the LR being investigated. If the ISLRN is assigned to the LR, the data center will be notified as well as the requester.

The back office will maintain a public list of LR and the associated ISLRNs so any interested party can check the state of assignments. The ISLRN server will not act as a substitute to catalogues of data centers and will not provide information (e.g. metadata) that are the responsibility of data centers and data right owners/providers.

An important feature of the server is to allow browsing of the database through ISLRNs as well as LR descriptions. It is essential that users are allowed to see the mapping between ISLRN(s) and LR(s).

5.3. Definition of a technical infrastructure

In practice, the executive team committee is to set up a technical infrastructure to implement the services (front and back offices) described above. Such infrastructure (will) comprises:

- A robust ISLRN server (phase 1).
- A deployed mirror server of ISLRN centers, in other geographical/linguistic areas (phase 2).
- A synchronisation process to share ISLRN between the various data centers and catalogue managing organizations (phase 2).
- A harvesting process to obtain information related to metadata from existing and agreed-with catalogues (phase 2).
- A Management interface of the back office for moderators and administrators (phase 1).

5.4. Management of the LRs

In order to keep the server useful for all users (while avoiding duplication of efforts with data centers and catalogues), the ISLRN server will also allow the update/addition of metadata to LRs that have been registered and assigned ISLRNs (phase 2), including information regarding LR versions.

6. ISLRN User Characteristics

6.1. User

It is not the (main) duty of this infrastructure to decide who is/are the right holder of a given LR and who can introduce a request for an ISLRN. The moderating team will do its best to check that the ”user” (who introduces such request) has some rights and that the metadata is informative enough to identify the stackholders.

A ”user” may simply browse the ISLRN server and look for a LR and its associated ISLRN. The following functionalities are available to him/her:

- Search of LRs that have been attributed an ISLRN and Mapping of ISLRN(s) and LR(s).
- Register as a provider to request ISLRNs for LRs (the registration is moderated).
- Contact a ISLRN administrator.
- Get information regarding ISLRN.

6.2. LR Provider

A LR provider is registered and owns credentials. However, “Log in” is not mandatory, it will be required only if one wants to submit a request and ask for an ISLRN assignment. The following functionalities are available to him/her:

- Log in with his/her credentials.
- Submit a form to query an ISLRN assignment.
- Look for metadata to identify LRs.
- Receive an ISLRN assigned to a submitted LR.
- Manage his/her account to update assigned LRs.
- Manage his/her account to check his/her ISLRN.
- Interact with the assignment team, including for appeals when a request is rejected.

6.3. Moderator

As indicated above, the whole procedure will be based on the executive committee that would provide a group of experts (moderators) to moderate the requests. Moderators will be loaded within the various institutions forming the executive team. A moderator, following an extensive investigation, accepts or rejects a LR submission for a new ISLRN. He/she moderates the queries in back-office with a restricted access. Credentials are given by an administrator to the moderator. At a basic level, a moderator can accept or reject a LR without any help. At a medium level, tools help the moderator to compare the LR in the database. At a more advanced level, the moderator is able to compare the submitted LR with other LRs from catalogues. The following functionalities are available to him/her:

- Log in to the back office with his/her credentials.
- Access accepted, rejected and pending LRs.
- Access the queries from providers.

- Check existing catalogues and sources of information about LRs
- Use the Internal forum to interact with the other experts/moderators.
- Contact the providers for further questions.
- Accept or reject the queries from providers.
- Modify LR assignment.

6.4. Administrator

An administrator is part of the executive committee IT services. His/her duties consist mostly in managing the ISLRN server and services. He/she is in charge of the maintenance of the server and the following functionalities are available to him/her:

- Log in to the back office with his/her credentials.
- Check the back up of the content of the database (included accepted, rejected and pending LRs).
- Add/remove a moderator to/from the moderator list.

7. Conclusion

This paper is a follow-up of the paper (Choukri et al., 2011) that introduced the principles of ISLRN and the arguments why we (the HLT community) should adopt a different identifier from those being used within e.g. the publishing industry (ISBN), the digital world (DOI), etc. In this paper we describe the modus operandi of the infrastructure to be established with its practical and technical aspects for assigning and using a unique identifier to LRs, namely International Standard Language Resource Number. It proposes the setting up of a networked and synchronised (mirrored) server to manage all these aspects and trusted to a small committee composed of some of the major international LR distribution and sharing institutions. Such committee could be set up under the supervision of a steering committee (the HLT Umbrella) composed of all active players, at the international level, within the computational linguistic and language technology field.

8. Acknowledgements

This work was partially supported by FlaReNet (ECP-2007-LANG-617001) and T4ME (META NET, FP7-ICP-4-249119).

9. References

Nicoletta Calzolari, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Irene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis. 2010. The lrec map of language resources and technologies. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Khalid Choukri, Jungyeul Park, Olivier Hamon, and Victoria Arranz. 2011. Proposal for the international standard language resource number. In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 75–83, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.