Language Resource Citation: the ISLRN Dissemination and Further Developments

Valérie Mapelli, Vladimir Popescu, Lin Liu, Khalid Choukri
ELDA/ELRA
9 rue des Cordelières, 75013 Paris, France
Email: {mapelli; popescu; lin; choukri}@elda.org

Abstract

This article presents the latest dissemination activities and technical developments that were carried out for the International Standard Language Resource Number (ISLRN) service. It also recalls the main principle and submission process for providers to obtain their 13-digit ISLRN identifier. Up to March 2016, 2100 Language Resources were allocated an ISLRN number, not only ELRA's and LDC's catalogued Language Resources, but also the ones from other important organisations like the Joint Research Centre (JRC) and the Resource Management Agency (RMA) who expressed their strong support to this initiative. In the research field, not only assigning a unique identification number is important, but also referring to a Language Resource as an object per se (like publications) has now become an obvious requirement. The ISLRN could also become an important parameter to be considered to compute a Language Resource Impact Factor (LRIF) in order to recognize the merits of the producers of Language Resources. Integrating the ISLRN number into a LR-oriented bibliographical reference is thus part of the objective. The idea is to make use of a BibTeX entry that would take into account Language Resources items, including ISLRN. The ISLRN being a requested field within the LREC 2016 submission, we expect that several other LRs will be allocated an ISLRN number by the conference date. With this expansion, this number aims to be a spreadly-used LR citation instrument within works referring to LRs.

Keywords: ISLRN, unique identifier, Language Resource citation

1. Principle and setting up

The International Standard Language Resource Number (ISLRN) was set up with the aim of providing a universal and unique identification schema, dedicated specifically to Language Resources (LRs) within the Human Language Technology (HLT) field, and under a free service for the HLT Community. This to ensure that LRs are correctly identified, and consequently, recognized with proper references for their usage in applications within R&D projects, product evaluation and benchmarking, as well as in documents and scientific papers.

After having reviewed a number of existing schemas (Choukri et al., 2011), the ISLRN identifier was implemented as a 12-digit random number as the new LR identifier, followed by 1 digit for a checksum number (see figure below).

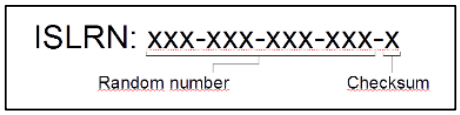


Figure 1: ISLRN identifier structure

In association to this identification number, a metadata schema was built in order to delegate semantics of the LR content to metadata which can easily and richly describe it, instead of integrating it within the number itself. Inspired by the broadly known OLAC schema, this metadata targeted on simple, easy and quick to fill in, fields to avoid any misunderstanding, and thus offers a minimal set of information describing the related LR, such as the name of the resource, the resource type, the source/URL where more information on the resource can be found, the description, versioning information, languages, etc.

ISLRN was endorsed by the NLP12[1] in 2013. In February 2014, ELRA (European Language Resources Association), LDC (Linguistic Data Consortium) and AFNLP/Oriental-COCOSDA announced the official opening of the ISLRN portal[2]. Details on the overall ISLRN infrastructure were already described in (Park et al., 2012).

2. Submission process

To obtain an ISLRN number of a given Language Resource, a producer, owner or distributor (whom we will name "provider" in this paper) of this LR has to register within the ISLRN portal. Once registered, this provider has the possibility to request an ISLRN by filling in the metadata presented as a one-page form. A provider has also the possibility to become a "certified provider", upon moderation, in order to be able to import one or more metadata descriptions in a pre-formatted way instead of filling in the description form by hand (XML format either from the ISLRN, the Meta-Share or OLAC schema).

Each submission to ISLRN follows a moderation process where the filled in metadata is checked carefully, mainly manually, by experts of the LRs field. For each LR submission, moderator checks a number of information, in particular:

- consistency of information with respect to the metadata fields, e.g. if the description field contains information that correspond to an actual description of a LR,
- consistency of information between the metadata fields, e.g. if the languages indicated in the Language field correspond to the description given in the Description field,
- accurateness of URL given in the Source field, i.e. an existing URL shall be indicated and information appearing in the resulting webpage shall be in relation with the described LR,
- duplication of information, i.e. if the LR was already submitted in ISLRN. To facilitate the work of identifying possible duplicates, a functionality was developed to enable the moderator to compare 2 or more LR submissions through a simple alignment of description forms.

[1] NLP12 is composed of the 12 Major Natural Language Processing and Computational Linguistics Organizations. The list can be found on the ELRA portal at http://ow.ly/LXwW4.

[2] http://www.islrn.org

From time to time, some interaction with the provider may be necessary to have the most consistent information possible before acceptance or rejection of the submission. The overall submission process is presented in Figure 2.

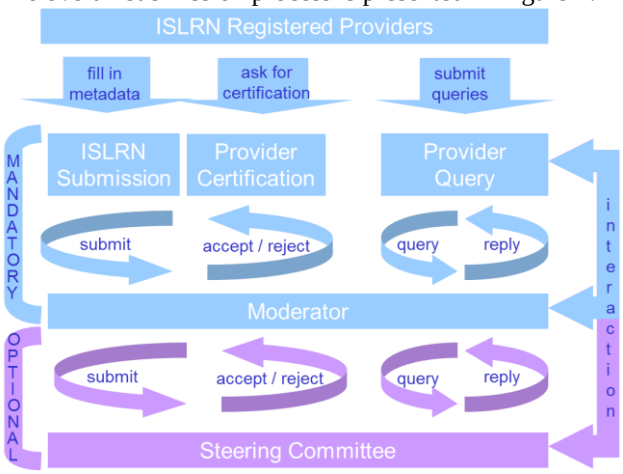


Figure 2: ISLRN submission process

When an LR has went through the whole moderation process and has been accepted by the moderator, an ISLRN is automatically associated to the resource and is notified to the provider. An example of an accepted LR is presented in Table 1 below.

Reference	NetDC Arabic BNSC (Broadcast News Speech Corpus)
Date of Submission	Jan. 24, 2014, 4:30 p.m.
Status	accepted
ISLRN	663-177-513-755-1
Resource Type	Primary Text
Media Type	Audio
Source	http://catalog.elra.info/product_info.php?products_id=13
Language	Arabic
Description	The NetDC Arabic BNSC (Broadcast News Speech Corpus) is a corpus developed by ELDA in the framework of the European-funded project Network of Data Centres (NetDC). The project was done in collaboration with the LDC (Linguistic Data Consortium) [...] A phonetic lexicon in Arabic SAMPA has also been included.
Version	1.0
Distributor	ELRA

Table 1: Example of a resource identified under ISLRN

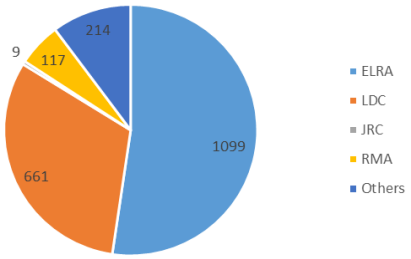
3. Dissemination

Since its launching, the ISLRN identification number has been adopted by large institutions over the world. As of the third quarter of 2016, 2100 LRs were allocated an ISLRN number. From the beginning, ELRA and LDC partnered to implement the ISLRN process and to assign the ISLRN 13-digit ID to all the LRs distributed in their respective catalogues. This is an ongoing process as all LRs being newly published in ELRA and LDC catalogue are constantly submitted to ISLRN. Anyone who is searching the ELRA and LDC catalogues can see that each LR is now identified by both the data centre ID and the ISLRN number. For the time being, over 1700 LRs obtained an ISLRN for ELRA and LDC combined catalogues.

When the portal opened to public, the Joint Research Centre (JRC)[3], the European Commission's in house science service, was the first organisation to adopt practically the ISLRN initiative by requesting ISLRN 13-digit unique identifiers to its LRs. Details on those LRs, including their ISLRN number, can be found on the Language Technology Resources pages of the JRC[4].

The Resource Management Agency (RMA)[5], an important LR player in South Africa, also adopted the ISLRN initiative and can apply for ISLRNs on behalf of the developers of the data that is managed and distributed via the RMA website. 117 language resources were submitted to the ISLRN, including language resources for the 11 official languages of South Africa. These include text and speech resources such as text corpora (annotated, genre classification, parallel), translation memories, custom dictionaries for government domain, compound semantic and splitting datasets, frequency word lists, speech corpora, and pronunciation dictionaries.

The graph below shows the number of LRs accepted in ISLRN distributed over the above mentioned organisations as well as the other submissions.



Graph 1: Distribution of ISLRN submissions

Thus, anyone who is using LRs available from those organisations, should now refer to the ISLRN number in their own publications.

4. Recent developments

Based on the web service exploitation experience, not only bugs were repaired but several developments were carried out. Such developments were implemented to facilitate the exploitation by providers of LRs, as well as moderators and administrators of the service.

The following new items were implemented:

- A specific URL was created to have a direct link to the metadata information for each LR described. Now, a user or a provider of an LR which has an ISLRN may not only refer to the ISLRN ID but also to the related metadata, by indicating the following URL: http://www.islrn.org/resources/XXX-XXX-XXX-XXX-X (XXX-XXX-XXX-XXX-X being the corresponding ISLRN number). For instance, the URL corresponding to the LR described in Table 1 is http://www.islrn.org/resources/663-177-513-755-1/.
- It happens that some LRs have several data providers/distributors and thus several URL where their metadata can be found. Previously, only one URL could be referred to. Consequently a multi-source URL was implemented to enable multi-pointers.

[3] https://ec.europa.eu/jrc/en/research-topic/internet-surveillance-systems

[4] https://ec.europa.eu/jrc/en/language-technologies

[5] http://rma.nwu.ac.za

- During the edition process of the metadata, multi-selection fields were presented in drop-down lists. However, some of those lists, like the selection of languages were too painful to fill in as such. For instance, the language list contains almost 8300 languages. To facilitate the finding of appropriate items, a search function including an autocompletion interface with user-friendly choices suggestion was implemented within the field.
- Although the XML submission in the ISLRN format is possible for certified providers, no information was provided on the format. Consequently, the ISLRN formalism was detailed in the webpage dedicated to the metadata[6].
- During the XML file submission, no error control was implemented. Now, a notification appears on the submission page and indicate which corrections must be done by the provider to submit the file correctly. For example, ISLRN languages consists of predefined and only readable information. The data flow in XML form coming from other platforms may use different types of language standards (ISLRN uses ISO codes). When a language submitted automatically did not correspond to the ISLRN language, it was simply rejected with no further notification. The automatic submission software module can now recognize other types of language codes different from the predefined ISLRN language codes, and it rejects the import with an error message notification to the submitter. When the updated importation module cannot find the language code in existing and predefined ISLRN languages codes, it imports all other information and creates an intermediate status ("incomplete") that allows the submitter to make modifications in the language field by finding and selecting the equivalent ISLRN language code. The correction module also offers a search suggestion and autocompletion interface as mentioned above.
- In order to deal more interactively with providers, an email management tool was built so that moderators and administrators of the service can exchange directly from the ISLRN web service. The message exchange module provides an on-line support service. A provider can submit his questions to a contact list that consists of "moderator" and "administrator" entities. Those questions are then submitted to the contact emails which are associated with the moderation and administration tasks, which can thus be a group of several persons from the ISLRN staff.

Beyond those developments specific to the ISLRN web service and due to the high number of LRs being published in data centres like ELRA or LDC, it was decided to implement an automatic submission API from a catalogue of LRs into the ISLRN system. This API was first developed for ELRA's catalogue and aims to be a first experiment for future integration within other catalogues like LDC's who already participates actively to this initiative. It enables to export the metadata extracted from the ELRA catalogue for all new resources and submit it directly into the ISLRN metadata. It is a first step towards fully automatic submission to ISLRN, with a minimal moderation and resource metadata matching for detecting duplicate submissions. For the time being, the moderation process remains the same, once the LRs are submitted within the ISLRN web service.

- Periodical statistics are caculated and displayed in the moderator's dashboard. In a chosen period of time, the table shows the number of resources per resources type, per media type, per language, per provider and the number of resources by date. For performance reasons, the queried results are cached for later enquiries, thus only new queries are calculated.

5. ISLRN, a key factor for accurate citation

In the research field, not only assigning a unique identification number is important, but also referring to a Language Resource as an object *per se* (like publications) has now become an obvious requirement. A study from Mariani & Francopoulo (2014) associated the use of a persistent identifier for Language Resources with an important parameter to be considered to compute a Language Resource Impact Factor (LRIF) in order to recognize the merits of the producers of Language Resources. As an example of citation requirement, we can mention the reference required within ELRA user agreements that state that any user of the Language Resources mentioned in those agreements "shall give appropriate references to Distributor, as well as to the name and reference of the Language Resources in scholarly literature when the Language Resources are mentioned. The following acknowledgement is required: "LANGUAGE RESOURCE NAME, ELRA catalogue (http://catalog.elra.info), ISLRN ID, ELRA ID".

The BibTeX standard[7] used to describe lists of references is a widely agreed standard for publications in the research field. The concept seems to be well adapted for citing Language Resources although the current model does not take into account Language Resources, nor ISLRN number. Thus, the idea is to make use of a BibTeX entry that could be used for bibliographical references and that would take into account Language Resources items. More work is currently ongoing in order to propose a template adapted to LRs.

BibTeX for a publication	**BibTeX for a LR**
@**Book**{hoel-71-whole, **author** = "Paul Gerhard Hoel", **title** = "Elementary Statistics", **publisher** = "Wiley", **year** = "1971", **series** = "Wiley series in probability and mathematical statistics", **address** = "New York, Chichester", **edition** = "3rd", **isbn** = "0-471-40300", }	@**LanguageResource**{Speecon, **author** = "Not often applicable but can be an option", **title** = "Dutch Speecon Database", **resourceType** = "Primary Text", **mediaType** = "Audio", **source** = " http://catalog.elra.info/product_info.php?products_id=1050", **publisher** = "Speecon Project, distributed via ELRA", **year** = "2014", **series** = "Speecon resources", **edition** = "1.0 (this is the version)", **islrn** = "613-489-674-355-0", }

Table 2: Comparison between a BibTeX for a publication and a BibTeX for a LR

As an example, the table above shows on the left-hand side an example of a BibTeX corresponding to a publication,

[6] http://www.islrn.org/basic_metadata

[7] http://www.bibtex.org/

whereas on the right is a template for a Language Resource, with specific fields dedicated to the description of a Language Resource. A stripped-down form of this template, without the *resourceType*, *mediaType* and *source* fields, is already being used for the LREC 2016 conference papers.

6. Further exploitation

The ISLRN being a requested field within the LREC 2016 submission, we expect that several other LRs will be allocated an ISLRN number by the conference date. With this expansion, this number aims to be a spreadly-used LR citation instrument within works referring to LRs.

As a second dissemination phase (see also Graph 2), more work aims to be carried out at different levels of the web service.

At the technical level, automatic submission to ISLRN from different large data catalogues should be carried out to facilitate their submission into ISLRN. This aims to be done for LDC's catalogue as a next step. This should be a path for other data centres or repositories like Meta-Share repository which already foresees the integration of the ISLRN ID field in its own metadata. The COCOSDA consortium also plans to add the ISLRN process within its future actions.

At the infrastructural level, the current management is held by ELRA with the strong participation of LDC. Still a committee should be constituted to manage and propagate the initiative as this was defined in the original plans. This could lead to 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 (see Figure 2). 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.

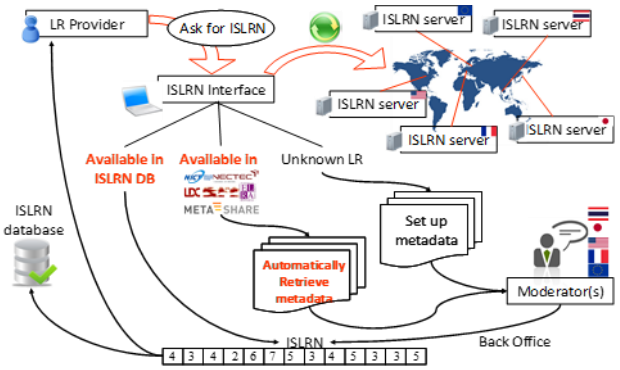


Figure 2: ISLRN attribution approach during a 2d phase

7. References

Jungyeul Park, Victoria Arranz, Olivier Hamon, and Khalid Choukri (2012). Using the International Standard Language Resource Number: Practical and Technical Aspects. In Proceedings of LREC'12. Istanbul, Turkey, 2012.

Khalid Choukri, Jungyeul Park, Olivier Hamon, Victoria Arranz (2011). Proposal for the International Standard Language Resource Number. In Proceedings of IJCNLP2011 Workshop on Language Resources, Technology and Services in the Sharing Paradigm, Chiang Mai, Thailand. November 8-13, 2011.

Joseph Mariani, Gil Francopoulo (2014). Language Matrices and a Language Resource Impact Factor. In Volume 48 of the series Text, Speech and Language Technology, Chapter "Language Production, Cognition, and the Lexicon". Springer, pp 441-471.