Proposal for the International Standard Language Resource Number

Khalid Choukri, Jungyeul Park, Olivier Hamon, Victoria Arranz
ELRA/ELDA
55-57, rue Brillat-Savarin
75013 Paris FRANCE
http://www.elda.org

Abstract

In this paper, we propose a new identifier scheme for Language Resources to provide Language Resources with unique names using a standardised nomenclature. This will also ensure Language Resources to be identified, and consequently to be recognised as proper references in activities within Human Language Technologies as well as in documents and scientific papers.

1 Introduction

Every object in the world requires a kind of identification to be correctly recognised. Traditional printed materials like books, for example, have generally used the International Standard Book Number (ISBN), the Library of Congress Control Number (LCCN), the Digital Object Identifier (DOI) and several other numeric 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 barcoding system for everyday products. Each of these schemes seems to have been the output of some specific need or circumstance within a domain.

In this paper, we review existing identifier schemes and conclude for the need to propose, specifically, the use of 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 standardised nomenclature. This will ensure that LRs are correctly identified, and consequently, recognised as proper references for their sharing 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 of 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 may be.

LRs imply corpora, dictionaries, and lexical and morphological resources in machine readable digital format. We also consider software tools for natural language processing and corpus-based computational linguistics as LRs if they can be stably packaged and deposited. They may include part-of-speech taggers, noun phrase chunkers, syntactic and semantic parsers, named entity recognisers, language modelling toolkits, corpus aligners, etc. Multimodal resources and systems also considered as LRs. Technology is in constant evolution and so are LR types, in their objective to help technological developments.

A citation has the purpose of acknowledging the relevance of the works of others. It attributes prior work to the original sources. It also allows the reader to provide a stable way of identifying proper references. However, the practice of using its proper identifier for LRs to cite and reference scientific data, along with individual resources as well as data sets, is less well developed (ISO-24619, 2011). LRs might be sometimes cited in a footnote even with several different names. For instance, the European Parliament Proceedings Parallel Corpus (Koehn, 2005) which is one of most cited LRs in the seventh International Conference on Language Resources and Evaluation (LREC2010),

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 75–83,
Chiang Mai, Thailand, November 12, 2011.

is cited by using several different names such as EUROPARL|EuroParl|Europarl (Parallel) (Corpus)[1]. In any case, a sad conclusion is that LRs remain in the background simply because the focus of the research is not on the resource per se (Calzolari et al., 2010).

The main goal for introducing the ISLRN for LRs is to get a unique way for naming a resource through the several LR distribution institutions. 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 centres that would manage their attribution. This is a mandatory step that will also have to work out the permanent localisation of a LR. The European Language Resources Association (ELRA) already has a role to discover, classify, collect, validate and produce LRs since 1995. Otherwise, the Linguistic Data Consortium (LDC), Gengo-Shigen-Kyokai (GSK), or Bavarian Archive for Speech Signals (BAS) play a similar role in the USA, Japan and Germany, respectively. However, current situation shows that each institution bears different types of identifiers even for the identical LR.

The remaining of this paper is organised as follows: We start by introducing a list of current identifiers in other domains (Section 2) and we also explore the actual LR identifiers introduced by several distribution institutions, in particular ELRA and LDC (Section 3). Then we explain the purpose of the new identifier for LRs and its associated metadata (Section 4). We provide our proposal for the new LR identifier (Section 5) and also provide previous other proposals for LR identifiers (Section 6), and we draw conclusions (Section 7).

2 Current Identification Schemes in Other Domains

Since we are forging a new identifier for LRs, we investigate in this section current identification schemes such as the ISBN for books, the AN in bioinformatics, the DOI and other schemes.

2.1 International Standard Book Number

The International Standard Book Number (ISBN) is used as a unique numeric book identifier. The 10-digit ISBN format was developed by the International Organization for Standardization (ISO) in 1970. Since 1st January 2007, ISBNs have contained 13 digits (See Figure 1). They consist of the EAN[2] code as GS1 prefix[3], the group identifier for language-sharing country group, the publisher code, the item number for the book title and a checksum character. The result is the ISNB such as 978-0060995058 for Milan Kundera's The Joke (English edition, published in 1993). Note that other than the check digit, no part of the ISBN will have a fixed number of digits[4]. For example, the group identifier can be from a 1- to 5-digit number such as 0 or 1 for English-speaking countries, 85 for Brazil, 99921 for Qatar, etc. In sum, ISBNs carry its own semantics derived from publishing industry practices.

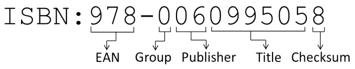

Figure 1: 13 digits ISBN.

2.2 Accession Number

An Accession Number (AN or AC) in bioinformatics is a unique identifier given to a Deoxyribonucleic acid (DNA) or protein sequence record to allow for tracking of different versions of that sequence record and the associated sequences over time in a single data repository. Researchers who wish to cite entries in their publications should always cite the first AN in the list (the primary AN) to ensure that readers can find the relevant data in a subsequent release. AN is used in several data resources such as the UniProt (SwissProt) Knowledgebase[5], GenBank[6], the EMBL Nucleotide Sequence Database[7], DNA Databank of Japan (DDBJ)[8], and Locus Reference

[1] That is, the corpus is cited as from simply EuroParl to more completely Europarl Parallel Corpus.

[2] EAN is for the International Article Number. Originally, it was the European Article Number.

[3] GS1 is an international association for the development and implementation of global standards such as the BarCodes identification system.

[4] http://www.isbn-international.org/en/manual.html

[5] http://www.uniprot.org

[6] http://www.ncbi.nlm.nih.gov/genbank

[7] http://www.ebi.ac.uk/embl

[8] http://www.ddbj.nig.ac.jp

Genomic[9], as identifier. While such sequence information repositories implement the concept of AN, it might have subtle variations. For instance, AN in the UniProt Knowledgebase consists of arbitrary 6 alphanumerical characters in the following format[10] (e.g. A1B123; P1B123; P12345):

1	2	3	4	5	6
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z,0-9]	[A-Z,0-9]	[0-9]
[O,P,Q]	[0-9]	[A-Z,0-9]	[A-Z,0-9]	[A-Z,0-9]	[0-9]

Entries can have more than one accession number when two or more entries are merged, or when an existing entry is split into two or more entries. However, AN has different syntax through data repositories which cannot provide an identical identification schemes

2.3 Digital object identifier

A Digital Object Identifier (DOI) is a unique identifier for digital documents and other content objects[11]. It provides a system for persistent and identification (Paskin, 2006). For example, a DOI name doi:10.1000/182[12], where 10., 1000 and 182 represent the DOI registry, the registrant, and item ID, respectively, can embed a URL using http://dx.doi.org and it is also linked as http://dx.doi.org/10.1000/182 which makes a DOI name actionable. In sum the DOI system (i) assigns a number which can include any existing identifier of any entity, (ii) creates a description of the entity associated with metadata, (iii) makes the identifier actionable which allows a DOI name to link to current data, and (iv) allows any business model in a social infrastructure. As claimed, DOI's Identifier is a network actionable identifier which means that "click on it and do something". It is irrelevant to LRs because some LRs may not have the referable site.

2.4 Other identifiers

Biomedical scientific research papers already have a PubMed IDentifier (PMID) which is a unique number assigned to each PubMed record[13]. PubMed is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals. PMID consists of arbitrary 8 digits. For example, a PMID 20011301 is for "Surgical management of locally advanced and locally recurrent colon cancer" (Landmann and Weiser, 2005)[14].

The canonical representation of an Electronic Product Code (EPC) is a Uniform Resource Identifier (URI) which is generally used to identify a name or a resource on the Internet. The EPC URI is a string having the following form[15]:

```
urn:epc:id:scheme:component1.component2...
```

where scheme names an EPC scheme. The precise forms of following parts such as component1, component2 depend on which EPC scheme is used. An example of a specific EPC URI is the following:

```
urn:epc:id:sgtin:0614141.112345.400
```

Each EPC scheme provides a namespace of identifiers that can be used to identify physical objects of a particular type[16].

2.5 Summary

Several identifiers have been described in this section which may be potential LR identifiers. Current identifier schemes are summarised in Table 1 with their name, an example for their syntax, their target object, their characteristics and relevance for the LR identifier. Since most of them are developed for other entities such as books for ISBN, DNA for AN, etc., they do not offer encoding schemes for necessary features for LRs. Therefore, we do not consider them relevant as LR identifiers. Moreover, ISBN is conceived especially for books and closely related to copyright law which may be different and complicated in each country. We do not believe the DOI name to be an optimal descriptor of a LR identifier, neither because of its actionable characteristic. As we mentioned, some LRs may not have the referable site as for various reasons, notably confidential company matters. On the other hand, since the DOI uses the Handle System, it is not for free.

3 Actual LR Identifiers

Most applications in Natural Language Processing (NLP) mainly depend on the existence of sufficient LRs regardless of their nature (raw data or annotated corpora). Several institutions for LR

[9] http://www.lrg-sequence.org

[10] http://www.uniprot.org/manual/accession_numbers

[11] http://www.doi.org

[12] This is an actual DOI number for *The DOI Handbook*.

[13] http://www.ncbi.nlm.nih.gov/pubmed

[14] http://www.ncbi.nlm.nih.gov/pubmed/20011301

[15] *EPCglobal Tag Data Standard* Version 1.5. See http://www.epcglobalinc.org

[16] *ibid.*

Name	Example	Target	Characteristic
ISBN	ISBN:978-0060995058	Books	Closely related to copyright law
AN	A1B123, AB123456	DNA or protein sequence record	Different syntax through data repositories
DOI	doi:10.1000/182	Digital documents and other content objects	assigned by the copyeditor
PMID	PMID17170002	Bibliographic database	Life sciences and biomedical information
EPC	urn:epc:id:sgtin: 0614141.112345.400	Every physical object	Limited to physical object

Table 1: Current Identifiers.

distribution in the world, in particular ELRA and LDC, have been responsible for providing a large part of the considerable amount of LRs in the domain. An increasing number of LRs are made available in catalogues. Currently, ELRA proposes two types of catalogue for LRs, the ELRA Catalogue[17] and the Universal Catalogue [18]. Similarly, the LDC's Catalog also provides hundreds of corpora and other language data[19].

3.1 Identifiers at ELRA

The ELRA Catalogue offers a repository of LRs made available through ELRA. The catalogue contains over 1,000 LRs in more than 25 languages. Other LRs identified all over the world, but not available through ELRA, can be also viewed in the Universal Catalogue. LRs at ELRA consist of spoken resources, written resources, evaluation packages, and multimodal/multimedia resources. Written resources also contain terminological resources and monolingual and multilingual lexicons. The actual LR identifiers in the ELRA Catalogue contain ELRA as publisher code, a systematic pattern (B|S|E|W|M|T|L) and 4 digits. B stands for a bundle which can contain several LRs within and S|E|W|M|T|L stand for Speech, Evaluation, Written, Multilingual corpora, Terminology and Lexicon, respectively. For example, the bundle package B0008 contains two separate spoken corpora: the LC-STAR Spanish phonetic lexicon (S0035) and the LC-STAR Catalan phonetic lexicon (S0048)[20]. While the ELRA Catalogue does not contain language processing tools as LRs at present, the Universal Catalogue does. Since ELRA is a partner of the Open Language Archives Community (OLAC), its Catalogue can be viewed as an OLAC repository[21], Oxford Text Archive[22], etc. Note that most of them only contain arbitrary digits as identifiers. ELRA is also sharing the index of its Catalogue through META-SHARE[23], a network of repositories developed within the META-NET network of excellence[24].

3.2 Identifiers at LDC

LDC assigns LDC as publisher code with a year number followed by (S|T|V|L) and 2 digits. S|T|V|L stand for speech, text, voice, and lexical(-related) corpora, respectively. The LDC Catalog is classified by data type and data source, or release year. LRs in the LDC Catalog are first divided into major categories according to the type of data they contain, and then are further broken down into minor categories based on the source of the data. For example, lexicon is further divided into dictionaries lexicon, field recordings lexicon, microphone speech lexicon, newswire lexicon, telephone conversations lexicon, varied lexicon and web collection lexicon. LDC also classifies software tools as LRs, such as LDC2004L01 for *Klex: Finite-State Lexical Transducer for Korean* (Han, 2004).[25]

3.3 Identifiers at other institutions

Among other institutions that are responsible for providing LRs, we explore identifiers at NICT, GSK, and BAS. The National Institute of Information and Communications Technology (NICT), and Nagoya University, for the purpose of developing LRs efficiently, have been constructing a large scale metadata database named SHACHI[26] as their joint project by collecting detailed meta-

[17] http://catalog.elra.info

[18] http://universal.elra.info

[19] http://www.ldc.upenn.edu/Catalog

[20] http://catalog.elra.info/product_info.php?products_id=980

[21] http://www.language-archives.org

[22] http://ota.ahds.ac.uk

[23] http://www.meta-net.eu/meta-share

[24] http://www.meta-net.eu

[25] Note that LDC also introduces the ISBN for LRs unlikely ELRA. For example, (Han, 2004) can be identified with the ISBN 1-58563-283-x as well as LDC2004L01.

[26] http://www.shachi.org

data information on LRs in Western and Asian countries (Tohyama et al., 2008). Identified LRs from other distribution institutions are assigned 6 unique digits by following C|D|G|T|N which represent corpus, dictionary, lexicon, thesaurus-like lexicon, terminology-related resources, and others respectively, as their own identifiers. For example, C-001543 is for Translanguage English Database (TED) where they crawl from LDC's LDC2002S04. Gengo-Shigen-Kyokai (GSK) (literally: 'Language Resources Association') was established in June of 2003 to promote the distribution of LRs in Japan.[27] The Language Resources Catalogue at GSK provides dictionaries and corpora. These are identified with 4 digits for the year and a capital letter chronically. For example, there are GSK2010-A for *Annotated Corpus of Iwanami Japanese Dictionary Fifth Edition 2004* and GSK2010-B for *Konan Kodomo corpus*. The Bavarian Archive for Speech Signals (BAS) was founded as a public institution in January 1995 and is hosted by the University of Munich, presently at the Institut für Phonetik und Sprachverarbeitung (IPS). BAS is dedicated to make databases of spoken German accessible in a well-structured form to the speech science community as well as to speech engineering[28]. They provide a set of Speech Corpora and Multimodal Corpora with acronym-style identifiers such as RVG-J for Regional Variants of German J which contains recordings of read and spontaneous speech by adolescents age 13-20[29]. Chinese-LDC (Chinese Linguistic Data Consortium)[30] assigns CLDC as publisher code, followed by a category, a 4-digit year code and a 3-digit identifier, for example, CLDC-SPC-2006-008 for a telephone speech recognition corpus. HLT-Centrale (Centrale voor Taal- en Spraaktechnologie, 'Dutch HLT Agency')[31] uses an acronym-style identifier per corpus, for example, 27MWC for a 27 Million Words Dutch Newspaper Corpus.

Table 2 summaries the types of identifiers used by those different institutions. Table 3 shows the number of LRs per institution by May 2011. To conclude, no identical LR has yet been formally identified through several institutions which leads same resource bearing two different identifiers. One such example is the Translanguage English Database (TED), which is catalogued both as ELRA-S0031 and LDC2002S04, that is, in two different ways. Our objective is to converge them using a unique way, that is, by forging a new LR identifier.

Catalogue	Number of LRs
ELRA	1,100+
LDC	500+
NICT	2,500+
GSK	10+
BAS	150+
Chinese LDC	90+
HLT-Centrale	50+
Universal Catalogue	1,800+
LRE Map	2,800+
Total (including duplicates)	9,000+

Table 3: Number of LRs of each institution.

4 Purpose of the New LR Identifier

4.1 Motivation

Identification of existing LRs is an essential, but a difficult and fastidious task. One has to find all available sources, from industry to university, from commercial to research. ELRA has promoted the collection and the dissemination of existing resources through its Universal Catalogue or more recently, the Language Resources and Evaluation (LRE) Map[32]. Both tools help to acquire knowledge using participative work. Another trend concerns the sharing of LRs through catalogues (see for instance, META-SHARE), where users (i.e. researchers, commercial users) are able to look for a large panel of data and tools. However, those two movements have shown several drawbacks which the community needs to take into account. One of them is linked to the nature of the LRs in the Internet era. Indeed, LRs have been created but also moved, duplicated, modified, or deleted. The consequence is that a LR may exist under various shapes, starting by its name, but also its format or even its content. Therefore, the community needs a unique way to identify, access, discover and disseminate LRs.

For instance, "Journal Officiel de la Communauté Européenne" and "JOC" refer to the same LR (ELRA-W0017). On the other hand, "Corpus EMILLE/CIIL" (ELRA-W0037) and "Corpus

[27] http://www.gsk.or.jp

[28] http://www.phonetik.uni-muenchen.de/Bas

[29] http://www.phonetik.uni-muenchen.de/forschung/Bas/BasRVG-Jeng.html

[30] http://www.chineseldc.org

[31] http://www.inl.nl/en/producten

[32] http://www.resourcebook.eu

	ELRA	LDC	NICT	GSK	BAS	Chinese LDC	HLT-Centrale
Publisher	X	X		X		X	
Category	X	X	X			X	
Year		X		X		X	
Digit ID	X (4)	X (2)	X (6)			X (3)	
Letter ID				X			
Free ID					X		X
Software	X	X		X	X		X
Example	ELRA-S0035	LDC2004L01	G-00035	GSK2010-C	SC10	CLDC-SPC-2007-002	CORN

Table 2: Summary of identifier designs per institution.

EMILLE Lancaster" (ELRA-W0038) are two different corpora and not just a different nomenclature for a same resource. It is about time that we are helped to refer to the LRs that we are using formally and clearly, without any risk of confusion or ambiguity. Accordingly, our goal is to allow the classification within catalogues, even redundant catalogues. For instance, the NICT catalogue contains mostly LRs from other catalogues, or OLAC get the export of LRs from many sources and necessarily duplicate inputs. The new LR identifiers that we want to propose, strictly granted, should avoid duplication of LR identifiers in the destination catalogues.

Actually, this proposal does not address the single issue related to LR catalogues, that is a desired way to share LRs. Another application of the identification lies in the production of documentation such as scientific papers or technical reports. Without the unique identification for LRs, we would struggle in the formal identification of any cited LRs within a document. LRs may be referred to by the new LR identification number instead of current usages such as URLs or author-invented names. This also overcomes the problem of wrong, broken or incomplete URLs.

A potential third application handles the tools and software that may use one or several LRs. Using a unique LR identification number eventually guarantees the correct use of LRs along with resource content and version. It is crucial that LRs should be used for evaluations without any bias. Our goal is then to define permanent localisations using the unique identifier for each LR used for HLT.

4.2 Metadata

Metadata schemas have been in constant evolution throughout the years. The non-stopping technological development makes it a requirement that its classifying or cataloguing procedures remain dynamic and open to the new arrivals in the field. Furthermore, different LR users have different needs, which can be observed both in the way the schemas are structured (from rather flat to very hierarchical) and the content of their components/elements, etc. (from rather limited to large and rich proposals). As it can be expected, the needs coming from LR providers or LR consumers range considerably. Likewise when we take into consideration the repositories themselves, with issues such as links, updating of information, etc. All this is being taken into consideration within one of the latest schemas still under development (the META-MD proposed within META-NET).

In order to name just a few of those different metadata schemas that have seen the light, we can refer to the Open Language Archives Community (OLAC)[33], which is Dublin Core-compliant, but only includes a small number of elements trying to prioritise interoperability over very rich descriptions. As already mentioned earlier in this paper, both ELRA Catalogue and Universal Catalogue, as well as the LDC Catalog provide very populated catalogues of LRs. Their metadata, although different, follows a 2-level hierarchy, covering LR types.

When it comes to identifying LRs, most metadata schemas have used different terms to refer to the resource names. However, as it has been mentioned in earlier sections, these names are not always consistent across catalogues, publications or other citations. Having a unique identifier that prevails beyond versioning and location changes, and that is unambiguous through LR searching and retrieving has also become a key issue for metadata. It is in this regard that the current proposal lies, with the creation of an unique identifier that will be registered within the metadata schema and

[33] http://www.language-archives.org

that will contribute considerably towards the life and sustainability of each resource implementing it. For such purpose, the metadata schema will contain an unique identifier element within its resource information component, and such element will allocate the standard identification number that the resource will have been assigned. Figure 2 depicts the idea behind this ID mapping, to show its "unique label" nature.

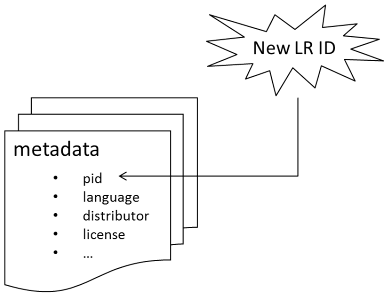

Figure 2: Mapping the new LR identifier to metadata as PID.

5 Proposal for the LR Identifier

In this section, a first formalisation of the International Standard Language Resources Number (ISLRN) is proposed. Then, several administrative characteristics that should be taken into account are defined.

5.1 Formal proposal for syntax

Such approach requires that an ontology in agreed upon within the community. Unfortunately, over the last couple of decades, no consensus emerged despite the number of proposals. It is easy to distinguish a large class of resources such as corpus versus lexicon, but within a corpus, we can imagine speech (signal and audio recordings) versus written texts. It is also difficult to build the commons over certain LRs such as a textual corpus consisting of transcribed audio data because one may always make a case that contradict such semantics. In this section, we review and criticise current practices for semantics of syntax introduced in current LR identification schemes.

- *Publisher* identifiers exist in ELRA, LDC, GSK and Chinese LDC classification. However, the ISLRN should not contain a publisher name, just as an institution name in general, because the distribution institutions are not usually a right holder of the LR and several institutions may distribute the same LR. An institution may also choose to distribute a LR anonymously.

- *Category* and *Type* identifiers are used by most of institutions. Even though it is important to keep an identification scheme symbolizing a categorisation, LRs can have very different categories and types as they evolve. Existing standards such as the BAMDES proposal (Parra et al., 2010) are also often limited, for instance it does not consider multimodal technologies. Moreover, the scope of LRs also leaves to LRs' provider and it make it more difficult to adopt proper categories or types.

- *Year* identifiers are used only by two institutions (LDC and GSK). Indeed, a resource may evolve over time and there may have a misunderstanding on the creation date, the delivery date or the last modification date.

- *Alphanumeric characters* identifiers are the most important, and are obviously used as identification schemes by all institutions, whatever they are digits or letters. Therefore, we should not avoid its introduction in the ISLRN. The size of the number should be decided according to the potential number of LRs (cf. Table 3).

One could suggest to add other semantics, but they are often limited to specific types of LRs. *Language* information, for instance, cannot apply to most of the multimodal technologies, and might not be easy when dealing with multilingual resources.

In sum, Publisher information do not appear in the LR identification scheme. As we mentioned before, wherever physical locations of LRs may be, a new LR identifier should be universal. A new LR identifier do not contain semantics about Category and Type, nor Year information. A LR identifier should delegate semantics of its syntax to metadata which can easily describe several semantics such as in DomainInfo, AnnotationInfo, etc., for example, in META-SHARE. Therefore, we decide to use 7-digit random numbers as the new LR identifier followed by 2-digit for version information and 1-digit for a checksum number. Having version information also allows us to describe LRs' granularity because information for resource bundles or resource collections can be encoded in Version information. The checksum number is encrypted from the preceding numeric identifier and version information. Our proposal is summarised in Figure 3.

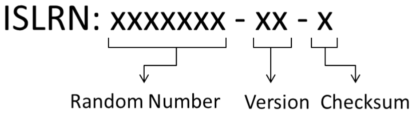

Figure 3: Proposal for the ISLRN syntax.

5.2 Administrative aspect

The definition of an ISLRN is certainly not the easier task, since administrative questions remain. First, the device to assign the ISLRN is crucial. ISLRN should be endorsed by major players and data centres, acting as an "umbrella" organisation. ISLRN attributions should be moderated, that is a small number of institutions should be granted the right to assign ISLRN. Prerequisite checking before assigning the ISLRN is also inevitable. LR Right holders or creators should provide minimum information to make their LRs be assigned ISLRN. Finally, we should pay attention to the legal issues regarding ISLRN and its usage. For instance, the ISBN is mandatory for printed, graphical and photographic documents subject ot a legal deposit. We may probably reflect the political importance of LRs as books are, meaning that the effort would be bigger than planned. However, the ISLRN should be assigned for free: no entry fee or no annual subscription: since the ISLRN will not be a legal deposit, the ISLRN is not an obligation, but rather an essential and best practice.

6 Other proposals for LR identifiers

FlaReNet (Fostering Language Resources Network)'s Blueprint of Actions and Infrastrucures would also "be a guideline for the LR community and National funding agencies, e.g. to prepare the ground for an EU directive concerning development of LRs at European scale"[34]. Currently, ISO already provides specifications both for the PID framework and its practice for referencing and citing LRs. The European Persistent Identifier Consortium also provides a service to name scientific data in a unique and timeless way.

6.1 ISO's PISA

Actually, ISO already proposed *Language resource management - Persistent identification and sustainable access (PISA)* as the International Standard (ISO-24619, 2011). It specifies requirements for the persistent identifier (PID) framework and for using PIDs as references and citations of LRs in documents as well as in LRs themselves (ibid.). It provides general guidelines for attributing PIDs for LRs as a part of a resource, a resource itself and a resource collection. The PID framework supports encoding of the PID as a Uniform Resource Identifier (URI), allows multiple URIs to render identifiers actionable without requiring client modifications, should be used to associated with metadata, and finally provides adequate security to change the PID-URI mapping or the associated metadata. ISO's PISA suggests Handle System (HS) and Archival Resource Key (ARK) as persistent identifier system implementations.

6.2 EPIC

The European Persistent Identifier Consortium (EPIC) provides a new methods to reference the scientific data in order to name in a universal way, which are permanent and citeable references.[35] It is not only for LRs, but for general scientific data. The Persistent Identifier Service is based on the Handle System like a DOI and uses as a prefix the number `11858`; the ordinary handle has the form `11858/flag-institution-num1-num2-num3-checksum` where its semantics explain themselves. Only flag is not defined yet and remains for special purposes such as derived handles.

6.3 Summary

While ISO's PISA has not provide concrete syntax for PID, nor other standardised techniques yet, EPIC explicitly introduces HS as PID system. As we mentioned before, there are LRs which may not have the referable site and the persistent identifier system cannot be applied. Therefore, previous proposals are not relevant to our purpose.

7 Conclusion

In this paper, we propose the ISLRN to provide LRs with unique names. This allows LRs to be identified, and consequently to be recognised as proper references. Therefore, the ISLRN can be

[34] `http://www.flarenet.eu`

[35] `http://www.pidconsortium.eu`

summarised as a unique identifier that allows to name and discover LRs. Actually, since we do not claim that the ISLRN is not a legal deposit, it is not an obligation. However, the ISLRN, when endorsed by major organisations involved in HLT, shall become an essential and best practice for LRs.

Acknowledgments

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

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 *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 19–21 May. European Language Resources Association (ELRA).

Na-Rae Han. 2004. *Klex: Finite-State Lexical Transducer for Korean*. Technical report, Linguistic Data Consortium, Philadelphia.

ISO-24619. 2011. *Language resource management – Persistent identification and sustainable access (PISA)*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, Phuket, Thailand, 12–16 September.

Ron G. Landmann and Martin R. Weiser. 2005. Surgical Management of Locally Advanced and Locally Recurrent Colon Cancer. *Clinics in Colon and Rectal Surgery*, 18(3):182–189.

Carla Parra, Marta Villegas, and Nria Bel. 2010. The basic metadata description (bamdes) and theharvestingday.eu: Towards sustainability and visibility of lrt. In *Proceedings of workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management at LREC 2010*, pages 49–53, Valletta, Malta, May. European Language Resources Association (ELRA).

Norman Paskin. 2006. *The DOI Handbook*. International DOI Foundation, Inc., Oxford, United Kingdom.

Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, and Hitoshi Isahara. 2008. Construction of an infrastructure for providing users with suitable language resources. In *Coling 2008: Companion volume: Posters*, pages 119–122, Manchester, UK, August. Coling 2008 Organizing Committee.