Henk van den Heuvel/Khalid Choukri

ELRA's central role in managing language resources in Europe

Abstract (English)

ELRA stands for European Language Resources Association. It is one of the key global players in managing language resources (LRs). In this contribution we focus on ELRA's position in the somewhat complicated landscape of LR providers, its services to its members, and its efforts to create a community of active and interconnected LR users, especially through the bi-annual International Conference on Language Resources and Evaluation (LREC). Furthermore, we highlight ELRA's legal expertise in licensing LRs both in academic and commercial contexts. Finally, we underline ELRA's sustainability as an LR intermediary, as demonstrated over more than 20 years of existence.

Abstract (Deutsch)

ELRA steht für "European Language Resources Association", das heißt "Euopäischer Verband für Sprachressourcen". ELRA ist einer der weltweit führenden Anbieter von Sprachressourcen (LRs). In diesem Beitrag beleuchten wir die Position von ELRA in der ziemlich komplizierten Landschaft der LR-Anbieter, wie auch die Dienste von ELRA für seine Mitglieder und das Bestreben, eine Gemeinschaft aktiver und vernetzter LR-Nutzer zu schaffen, insbesondere durch die alle zwei Jahre stattfindende International Conference on Language Resources and Evaluation (LREC). Darüber hinaus betonen wir die juristische Expertise von ELRA bei der Lizenzierung von LRs sowohl in akademischem wie auch in kommerziellem Kontext. Schließlich unterstreichen wir die Nachhaltigkeit von ELRA als LR-Vermittler, welche sich in den mehr als 20 Jahren seines Bestehens bewährt hat.

1. What is ELRA?

Language Resources are at the centre of ELRA's activities and are its reason for existing. So let's first define what language resources are. The term Language Resources (LRs) refers to sets of language data and descriptions in machine readable form. These are used in many different areas: components, systems, applications (including the creation and evaluation of natural language, speech and multimodal algorithms and systems), software localisation and language services, language-enabled information and communication services and knowledge management (e-commerce, e-publishing, e-learning, e-government). ELRA has extended the concept of LRs to include videos, images and other multimedia and interaction modalities.

ELRA, which stands for the European Language Resources Association, is one of the key players in managing LRs. It was created in February 1995 with initial funding from the European Commission. ELRA is a not-for-profit association of users of Language Resources (LRs). The main reason for creating ELRA was (and still is) to highlight the need for the mutual exchange and re-use of LRs. Through its executive organisation ELDA,[1] ELRA has established itself as a repository centre responsible for all aspects of LRs, including:

– Discovery, sharing and distribution for R&D and commercial purposes;
– Archiving/cataloguing, licensing, production, quality assurance and sustainability analysis.

Thus, ELRA is a broker /middleman for LRs. For its academic and commercial members it helps with IPR & licensing issues in order that they can distribute their own LRs and obtain those of others.

In this paper we will highlight ELRA's position in the somewhat complicated landscape of LR providers, its services to its members, its mission to create a community of active and interconnected LR users (including a description of its forum, the LREC series of international conferences), and its sustainability as an LR intermediary, evidenced by its 20-plus years in existence.

But first let's have a closer look at ELRA's members, services and activities.

2. ELRA Members

ELRA's members come from both academia and business. ELRA is Europe-oriented and the majority of its members are European organisations. However its resources also attract non-European interest, as Table 1 shows.

2016	**Academia** (not-for-profit org.)	**Industry** (profit-making org.)
Europe	23	11
Non-Europe	11	4

Table 1: Distribution of ELRA members in 2016

To date, ELRA's membership scheme has been institution-based. Any organisation, public or private, European or non-European, can join. However, membership with voting rights is available only to organisations legally established in Europe, in accordance with article 5 of ELRA's statutes. Changes are under way to provide individual membership within a specific college of members.

[1] http://elda.fr/en/.

Figure 1 displays the distribution and longitudinal flow of members. A decline in members over the last few years is clearly visible. This is one of the reasons why ELRA has introduced membership for individuals from 2018; this also supports ELRA's mission to connect LR users worldwide.

Fig. 1: ELRA's membership distribution and evolution at a glance

3. ELRA's services and activities

What does ELRA offer to its members? Here is an overview of its services:

- **Price discounts** on resources, with a reduction of up to 70% on some items, including the resources produced by ELRA.
- **Discounted fees** for members registering for the biennial conference LREC.
- **Legal and contractual assistance** are also provided to the members of the association.
- **Regular information** through the monthly members' news bulletin and updates on the www.elra.info web site.
- Free online access to the **Language Resources and Evaluation (LRE) Journal**, published by Springer, is provided to all ELRA members.

All services are available to both institutional and individual members. The only exception is the discount on LRs, which can only be sold to institutional members due to legal/contractual reasons. However, ELRA also has rights to a number of LRs with free distribution and these are available to all its members. Moreover, many of the resources in ELRA's catalogues are freely available for research purposes, and can therefore also be obtained under a free license.

Last but not least, ELRA is the initiator and organiser of the International Conference on Language Resources and Evaluation (LREC),[2] which is the major bi-annual conference on LRs. LREC is an international scientific event which

[2] See http://elra.info/en/lrec/ and http://www.lrec-conf.org/.

aims to provide an overview of the state of the art, exploring new R&D directions and emerging trends, exchanging information regarding LRs and their applications, evaluating methodologies and tools, on-going and planned activities, industrial uses and needs, requirements coming from e-society, with respect to policy, technological and organisational issues.

LREC provides a unique forum for researchers, industrial players and funding agencies from across a wide spectrum of areas to discuss problems and opportunities, to find new synergies and to promote initiatives for international cooperation in support of research into language sciences, progress in language technologies, and the development of corresponding products, services, applications, and standards.

Over the last 20 years LREC, which takes place every other year, has become the major event on Language Resources and Evaluation for Human Language Technologies. The conference programme is organised around parallel oral and poster sessions during the main conference (3 days) and workshops and tutorials (2 days before and 1 day after the main conference). Since 1998 the success of the conference has grown, and it attracted over 1,200 participants and around 750 paper presentations in 2016 (oral and poster presentations).

Figure 2 gives an overview of ELRA's activities. The figure clearly illustrates ELRA's central position between LR providers and LR users. ELRA identifies LRs that can be shared and negotiates with the providers on appropriate licenses and prices (for paid-for resources). The LRs offered are also validated before they enter ELRA's catalogue. The validation manuals, including what is known as the Quick Quality Check (QQC) are available to the community (and are being packaged as part of the ELRA Data Management Plan, ELRA-DMP).

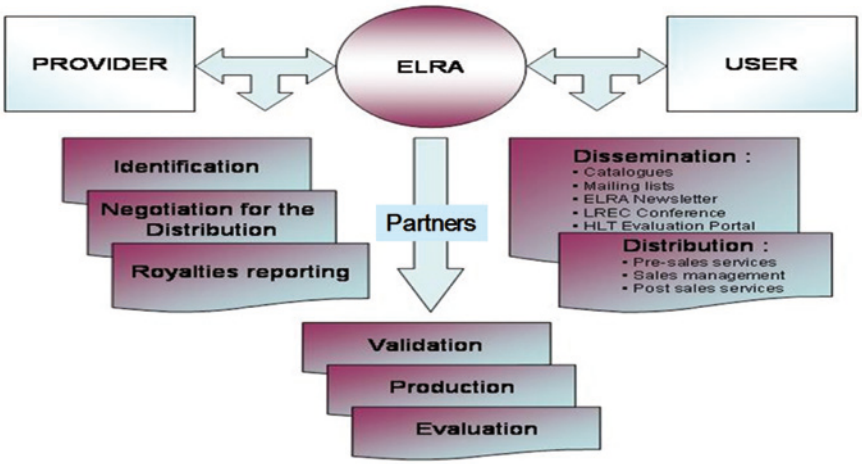

Fig. 2: Overview of ELRA's activities

Upon successful completion of the validation process, LRs are notified to users through various means: catalogues on ELRA's website, the newsletter for members and various portals.

Further information about ELRA's activities and services can be found in Choukri et al. (2016).

4. ELRA's position in the LR management landscape

The area of LR brokers and intermediaries is currently somewhat confusing. It is not easy to clearly delineate the various players in such a way that their activities and target users can be easily distinguished. In our opinion there are three relevant dimensions to be picked out:

1) European/American,
2) Tools/data,
3) Commercial/non-commercial.

In what follows we will position ELRA between two other major players: LDC[3] and CLARIN.[4] LDC is the Linguistic Data Consortium, based in the USA and managed by the University of Pennsylvania. Unlike ELRA, LDC's production of resources is constant and on demand (e.g. DARPA) while – although ELRA produces many LRs – these are often done as part of customer projects. In addition, there are clear differences in membership policies between ELRA and LDC. LDC offers resources per membership year at a higher membership fee, with free resources for members during the year. ELRA has a considerably lower membership fee with discounts on resources (rather than offering all of them for free for the membership period). The difference in policy is clearly attributable to the fact that LDC can produce most of its own resources on a yearly basis as indicated above.

It should be stressed that ELRA and LDC cooperate closely in offering joint resources, organising LREC, and adopting ISLRN[5] as a unique identifier for LRs.

We also see clear similarities and differences between ELRA and CLARIN. CLARIN and ELRA both serve the Social Sciences and the Humanities (SSH) and the language engineering community in offering services around language resources (LRs). CLARIN ERIC and ELRA have their own background and position in this landscape. Their activities are both overlapping and complementary. The organisations overlap in:

– The identification and dissemination of language resources,
– Expertise in multilinguality and in Language and Speech Technology (LST),
– European identity.

[3] https://www.ldc.upenn.edu/.

[4] https://www.clarin.eu/.

[5] http://elra.info/en/islrn/.

The organisations are complementary in:

– Their target users: both address academic and commercial users, but CLARIN has a stronger connection to academia and ELRA to industry, for which it has business models.
– Resources: ELRA is a distribution centre for data, including evaluation data, while CLARIN gives access to data and tools.
– ELRA focusses mainly on LST for Research and Development whereas CLARIN has a primary focus on support for research in SSH and LST.[6]

CLARIN and ELRA see clear advantages in collaboration. Collaboration helps ELRA to strengthen its position in academia (an explicit aim in its new membership policy, which introduces individual membership and related services). Collaboration helps CLARIN to make stronger connections with industry, which is instrumental in widening the audience for the resources it offers, including matching and training young researchers in the CLARIN community.

5. ELRA's Language Resources

Some of the following passages from ELRA's 2016 Annual Report give an idea of the number and types of LRs that ELRA offers through its catalogues.

The basic distinction is between Speech LRs, Written LRs and Multimedia LRs. For the Written LRs we distinguish lexica of various sorts, corpora and terminology databases.

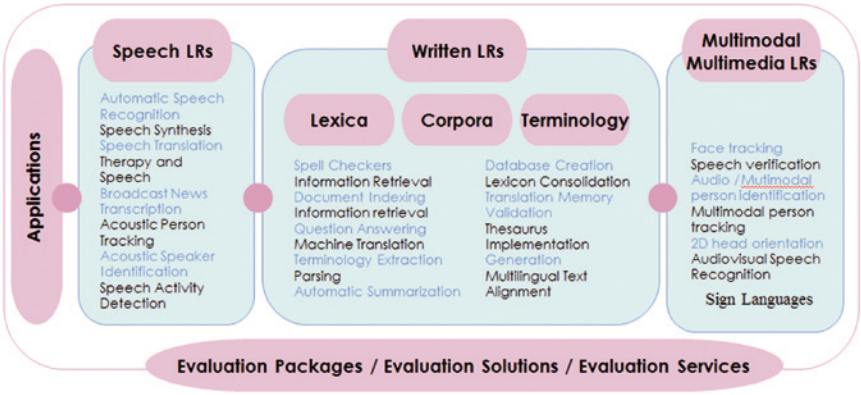

Fig. 3: LR types offered by ELRA

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

Our LRs are suited for research purposes and for building practical and commercial applications. ELRA and ELDA have a long tradition of supplying evaluation packages for targeting performance evaluation and benchmarking in R&D.

Figure 3 shows a more detailed overview of the types of LRs that are distributed through ELRA.

Currently ELRA has 1,155 LRs in its catalogue, as shown in Figure 4, which presents a graphical view of the number and types of LRs offered via various catalogues. There are three catalogues worth noting.

Fig. 4: Overview of catalogue and distribution figures

5.1 Language Resources Catalogue

The Language Resources collected by ELRA are made available to the public through the ELRA Catalogue, accessible online at http://catalog.elra.info. The distribution of resources, which cover a wide variety of languages and belong to different modalities, is shown in Figures 3 and 4.

Early in 2018 the catalogue was completely redesigned, with a new interface and improved navigation. The new catalogue allows visitors easier access to the 1,075 Language Resources (LRs) and their corresponding descriptions. Among

the new features, the catalogue now offers an extended metadata to describe the LRs and a refined search of the catalogue data in order to find more specific information using criteria such as language, resource or media type, license, etc.

Currently, LRs can be selected and placed in a cart, from where the user can send a request for a quotation to initiate the order. When logging in, the user selects LRs and obtains distribution details (licensing information, prices) depending on his/her user status: ELRA member/non-member, research vs commercial organisation. Full e-commerce integration will be completed at a later stage.

More functionalities pertaining to the ELRA Catalogue, including the ISLRN automatic submission and the e-licensing module (automatic filling in and electronic signature), will also be developed and integrated.

5.2 Universal Catalogue

The Universal Catalogue is an important identification feature. Information regarding Language Resources (LRs) identified all over the world is gathered in this publicly available repository. The LRs are generally located by the ELRA team, but external feedback from our members, collaborators, and website visitors is also included. Our aim is to provide researchers and developers with information about existing LRs and spare them the effort of searching or rebuilding similar resources. The Universal Catalogue is collaborative: through a simple webform, anyone can add some basic information, point to an existing language resource or enrich the current description of an LR already present in the Universal Catalogue.

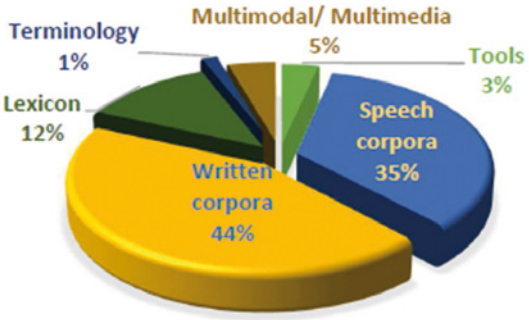

Fig. 5: Distribution of LRs in the Universal Catalogue

The Universal Catalogue is accessible through the following web site: http//. universal.elra.info.

5.3 LRE Map

Initiated by ELRA and FlareNet at LREC 2010, the LRE Map is a mechanism intended to monitor the use and creation of language resources by collecting information on both existing and newly-created resources during the submission process. The feature has been so successful that it has also been implemented at other major conferences including COLING, IJCNLP, Interspeech, LTC, ACLHT, O-COCOSDA and RANLP, in addition to the *LRE Journal.*

The LRE Map feature is now part of the LREC standard submission processes for both the main conference and all the workshops. For LREC 2014, 1,070 LR forms were filled in (compared with 900 in 2012). Globally, over 6,000 LR resource type forms have been completed at all the conferences, including LREC 2016, which have adopted the LRE Map.

At LREC 2014, as a new feature, the LRE Map was offered to authors so that they could share their resources by uploading them during the submission process.

All information on the LRE-Map can be found at http://lremap.elra.info/.

6. Legal Support

Using, producing, sharing or distributing Language Resources can trigger legal questions related to Intellectual Property Rights (IPR) management which are not always easy to resolve. For nearly 20 years now, ELRA has established close co-operation with legal experts to clarify such IPR issues, to design licensing schemas and draft licenses, but also to provide assistance on any contractual or legal matter that may arise during the Language Resources life cycle, including the acquisition, production, sharing, or distribution phases.

With this Helpdesk service fully dedicated to IPR issues, we are extending our legal support to the whole Human Language Community. This Helpdesk provides services similar to those offered by the META-SHARE network.

ELRA's expertise can be of great value for academics who would like to share relevant LRs. ELRA provides a legal framework for making the LRs available for both academic research (free of charge) and for business (at a commercial price).

To facilitate an easier understanding of the various licenses that exist for the use of Language Resources (ELRA, META-SHARE, Creative Commons, etc.), ELRA has developed a License Wizard[7] to help rights-holders to share/distribute their resources under the appropriate license. It is also available to users so that they can better understand the legal obligations that apply in various licensing situations.

[7] http://wizard.elra.info/.

The License Wizard works as a web configurator that helps the user to:

– select a number of legal features;
– obtain the user license adapted to their selection;
– define which user licenses they would like to select in order to distribute their Language Resources;
– integrate the user license terms into a Distribution Agreement that could be proposed to ELRA or META-SHARE for further distribution through the ELRA Catalogue of Language Resources.

7. ELRA and sustainability

When it comes to sustainability ELRA has a longstanding track record as an LR intermediary, proven over more than 20 years of existence. Sustainability issues cover a wide range of aspects related to LR management. Within FLaReNet,[8] ELRA has identified twenty relevant factors impacting sustainability:

1) LR specifications (including references to best practices and standards),
2) Production and management of LR documentation,
3) Quality assessment and quality validation report,
4) Management of rights, ethics, privacy, consent, and other sensitive legal issues,
5) Information dissemination including scientific publications,
6) LR format, encoding, content → interoperability,
7) LR portability across languages, environments and domains,
8) LR packaging (compilation of items, including resource documentation),
9) Rights to be granted and licenses,
10) Data identification, metadata and LR discovery,
11) Versioning and referencing of the LR (ISLRN),
12) Usability assessment and relevance,
13) Accessibility of the LR (LR package, medium),
14) Accessibility of LRs in an "open" mode ("open" here means open data),
15) Preservation of LR media for long-term access,
16) LR access charge (LR for free /for a fee),
17) Reference to production and use projects, environments,
18) Relevance for other NLP applications and areas,
19) Maintenance and support over time,
20) Role and impact of data centres and archiving houses.

These factors have well been taken care of by ELRA. This list offers a relevant checklist for any LR developed at the various research institutes and language councils that are part of the EFNIL community. ELRA is happy to offer further assistance to each interested EFNIL member.

[8] http://www.flarenet.eu/.

References

Choukri, K./Mapelli, V./Mazo, H./Popescu, V. (2016): ELRA activities and services. In: *Proceedings of the 10th International Conference on Resources and Evaluation, May 23-28, 2016, Portorož, Slovenia*. Portorož: ELRA, 463-468. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1250_Paper.pdf.

Choukri, K./Van den Heuvel, H. (2017): ELRA Annual Report 2016. Paris: European Language Resources Association (ELRA). http://elra.info/media/filer_public/2017/09/12/annual_report2016body-vf3.pdf (last visited on 9 Feb. 2018).

ELRA website: http://elra.info/en/.

Bibliographical information

This text was first published in the book:
Gerhard Stickel (ed.) (2018): National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences. [299 pages.]

The electronic PDF version of the text is accessible through the EFNIL website at:
http://www.efnil.org