Recent Developments within the European Language Resources Association (ELRA)

Khalid Choukri, Audrey Mance, Valérie Mapelli

European Language Resources Association (ELRA) &
European Language resource - Distribution Agency (ELDA)
55-57, rue Brillat-Savarin
75013 Paris France
{choukri, mance, mapelli}@elda.fr

Abstract

The main achievement of ELRA (the most visible) is the growth of its catalogue. The ELRA catalogue as of April 2000 lists 111 speech resources, 50 monolingual lexica, 113 multilingual lexica, 24 written corpora and 275 terminological databases. However, many Language Resources (LRs) need to be identified and/or produced. To this effect, ELRA is active in promoting and funding the co-production of new LRs through several calls for proposals. As for the validity of the existence of ELRA for the distribution of language resources, the statistics from the past two years speak for themselves. The 1999 fiscal report showed a rise with the sale of 217 LRs (122 for research and 95 for commercial purposes; with speech databases representing nearly 45%), compared to the sale of 180 LRs (90 for research and 90 for commercial purposes; with speech databases representing nearly 65%), in 1998 and to 33 sold in 1997. The other visible action of ELRA is its membership drive: since its foundation, ELRA has attracted an increasing number of members (from 63 in 1995 to 95 in 1999). This article is updated from a paper presented at Eurospeech'99.

1. Introduction

ELRA is a useful conduit for the distribution of speech, written and terminology databases, enabling key players to have access to LRs. In order to effectively produce and provide such resources to research and development groups in academic, commercial and industrial environments, it is necessary to address technical, commercial, legal, logistic and other practical issues. This has already been done by ELRA through the establishment of an operational infrastructure that capitalizes on the investments of the European Commission and other European National agencies to ensure the availability of Speech, Text, and Terminology resources.

After five years of activity, ELRA has managed to make available, worldwide, a large set of marketable resources. ELRA has handled the legal issues through generic license agreements that are made widely available. A set of validation manuals have been produced and widely distributed.

It is of paramount importance that ELRA reaches fruitful agreements with other regional organizations in order to achieve altogether a better streamlining of efforts in the development of new Language Resources that are of interest to "global" players. A joint initiative to start a cooperation between ELRA and the US Linguistic Data Consortium (LDC) has been recently launched. Fruitful contacts are forged with Oriental Cocosda as well as with GSK (Gengo Shigen Kyouyuukikou) for Asia and Japan.

As for the validity of the existence of ELRA for the distribution of language resources, the sales statistics from the past three years speak for themselves.

The 1998 fiscal report, covering the period October 97-September 98, indicated a total of 180 sales whereas the 1999 fiscal report, covering the period October 98-December 99 (in order to shift the ELRA fiscal year to the calendar year, the 1999 management report covered the fiscal year period October 98 – December 99), showed an increase with a total of 217 sales, 122 of which were for research and 95 were for commercial purposes. Of these 217 purchases, 97 were specifically speech databases, representing nearly 45% of the yearly sales. Legal and contractual assistance is also included in the services for members, as ELRA also seeks to simplify the relationship between providers and users (customers) by drafting generic agreements that specify the responsibilities and duties of each party when licensing a language resource. Since October 1996, 92 agreements with providers of language resources have been secured by ELRA, preventing users and providers from spending their time on contractual agreements negotiations. In order to add value to the resources it distributes, ELRA also initiated the production of validation manuals for each resource type, namely spoken resources, written corpora, and lexica.

The highlight of 1999-2000 for the association is the organization of this Second International Conference on Language Resources and Evaluation (LREC), Athens, Greece. It was initiated by ELRA and organized in co-operation with other national and international associations and consortia.

2. The Association

Any legally registered organization can join the association, though full membership, with voting rights, is available only to organizations legally established in Europe. For 2000, the annual membership fee is as follows:

Non-profit making organizations	750 EURO
European small/medium-sized companies < 50 employees	1000 EURO
European profit making organizations >= 50 employees	1500 EURO
Non-European profit making organizations	5000 EURO

Table 1. Membership fees

The difference between membership fees of European and Non-European companies reflects the grants received from the European agencies.

Since its foundation, ELRA has attracted an important and steady number of members as shown below:

1995	1996	1997	1998	1999
63	**70**	**75**	**81**	**95**

Table 2. Number of ELRA members

Throughout the last three years we have noticed a global steady membership base. If we consider the sectors of activities (speech, text, terminology), we notice a particular decrease in the terminology sector and a relatively important increase in the speech sector. This year we had about 95 members. We can also point out that out of the 95 members of 1999, 23 joined ELRA in 1999.

The services offered by ELRA to its members are summarized both on the ELRA web site (see http://www.elda.fr) and in brochures. These services go beyond the important discount given on the price of language resources.

3. ELRA's mission and technical activities

As stated above, ELRA has been entrusted with a crucial mission: to ensure that Language Resources needed by Language Engineering players are made available when they already exist or to produce them in a cost-effective framework. This mission is tuned from time to time to anticipate future requirements. Such a mission can be itemized as: identification of useful resources, handling the legal issues related to the availability of Language Resources, distribution activities and pricing policy, validation and quality assessment, commissioning the production of needed Language Resources, market watch, and information dissemination.

3.1 Identification of Useful Resources

In order to play its role, ELRA committed to create structured and publicly available catalogues of Language Resources. In order to do so, ELRA has prepared a set of description forms to help the providers describe what they propose to ELRA for distribution in a more uniform and consistent way (see the URL corresponding to spoken databases at: ftp://ftp.icp.grenet.fr/pub/elra/spee-en.ps). This form includes all the features one needs to know about a speech database: file standards, acquisition conditions (Telephone vs Microphone, environment), annotation levels, etc. Other forms (lexica, corpora) are available from the same site.

In the first catalogue, the identified resources were at different status levels: available through ELRA, available through the owner, under negotiation but not available yet, available on a case by case basis, identified but not available, etc. There were simply too many categories which made things too complex to be understandable.

Since then we have been committed to publishing a catalogue of resources that are available via ELRA or, in a very few cases and for very sensitive databases, available through the owner/provider.

Our catalogue is compiled with respect to the three colleges of ELRA: speech, written, and terminology. Some tools can be also catalogued if they are available for free.

The progress of our identification task is illustrated on the following histogram:

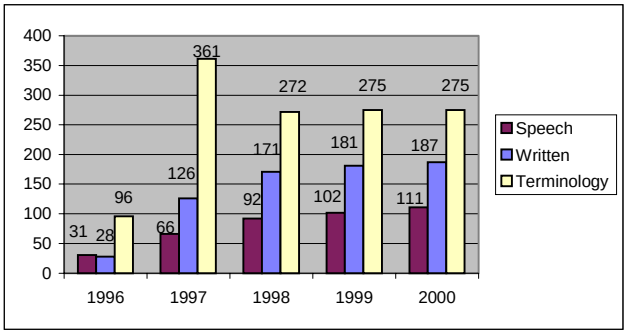

Figure 2. Histogram of the ELRA Catalogue of resources

(Some resources have been deleted from our catalogue due to negative quality assessment and some others due to various mergers of providers which led to the termination of some distribution contracts).

In the speech area, we can see that the catalogue has grown from the 22 initial resources of March 96 to about a hundred today. This does not hide that many key resources are still not available for a large number of languages (including Western European ones). If we consider some basic resources that should be available for all languages, such as:

- Articulatory databases (e.g. ACCOR),
- Basic speech data with some phonetic material and some phonetic sequences, by a small number of speakers, recorded in a quiet environment (e.g. EUROM-1 & BABEL),
- Pronunciation lexicon (e.g. BDLEX, PHONOLEX),
- Proper names pronunciation lexicon (e.g. ONOMASTICA),
- Newspaper read text (e.g. BREF, Siemens-100, Apasci),
- Basic telephone speech (e.g. SPEECHDAT),
- Telephone-based speaker verification (e.g. PolyVar),
- Text corpora for language models (e.g. MLCC, Le Monde …),

and if we consider Western European languages, our catalogue can be represented by the table given in Table 3.

Speech	UK	I	F	SF	G	EI	SP	Pt	Gr	NL	Dan	Sw	Finn	Nor
Articulatory database	A	E	E		E	E						E		
Basic speech data	A	A	A		A	E	E	E	E	E	E	E		E
Pronunciation lexicon		A	A		A					A				
Proper names pron. lex.	U E	E	E		A***			E		E	E	E		
Newspaper read text		A	A		A		E	E		A				
Basic telephone speech	A	A	A	A	A	U	A	A	E	A	A	A	A	A
Teleph. Speaker verif.		A	E	A										
text corpora for language Models	A	A	A		A		A			A				

Table 3. Availability of Language Resources per language
(UK: British English, I: Italy, F: French, SF: Swiss French, G: German, EI: Irish, SP: Spanish, Pt: Portuguese; Gr: Greek, NL: Dutch, Dan: Danish, Sw: Swedish, Finn: Finnish, Nor: Norwegian)
A: Available through ELRA;
S: Available through ELRA within the next quarter;
E: Exist/identified but not (never!) available;
" " (blank): Probably Not available / has not been identified;
U: Under completion/Well advanced project with distribution plans
*** Available through German telecoms

This matrix illustrates that many basic resources are not available and that there is a need to stimulate their production in order to meet the needs and requirements of both academic institutions and industrial users. A survey of national programs recently initiated by ELRA and other partners will allow to have more up-to-date information on this topic. Part of the survey aims at filling such matrix for various areas (written corpora, lexica, multimodal databases, etc.), while extending it to all European countries and languages.

3.2 Handling the legal issues

The basic principles of language resource licensing has been worked out with the support of lawyers. At the beginning, marketing Language Resources was a new activity, and creating an equitable and balanced framework was not easy. It was agreed that one of the priority tasks of ELRA was to simplify the relationship between producers/providers and users of LRs.

In order to encourage producers and/or providers of LRs to make such data available to others, ELRA has drafted generic contracts defining the responsibilities and obligations of both parties.

To minimize variations in agreements and to keep things simple, these contracts are based on the following model:

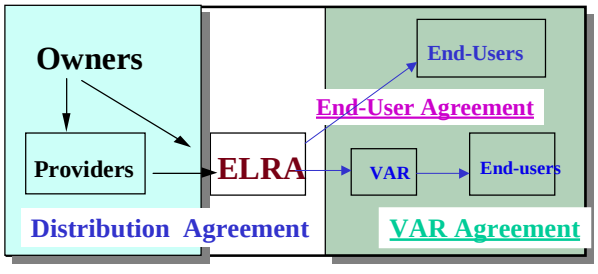

Figure 1. ELRA Agreement model

This model takes into account the interest of both parties (producers and users) in keeping with ELRA's role as a neutral, non-profit organization, dedicated to promoting the language engineering field. Contracts (or Licenses) are drawn up between ELRA and the resource provider and/or ELRA and the resource user (either a VAR, Value Added Reseller, or End-user). Since 1996, they have evolved in the light of feedback from our members, customers and resource providers.

Let us illustrate the role of ELRA with our resource referenced as ELRA-W0023, the first set of which consists of 6 written corpora of similar nature, provided by six different newspapers through Europe (Le Monde from France, Financial Times from UK, Handelsblatt from Germany, Expansion from Spain, Il Sole 24 Ore from Italy, and Het Financieele Dagblad from The Netherlands). ELRA has signed one contract with each provider. If this resource is purchased via ELRA, the customer needs to sign one agreement. If the customer rather chooses to go to each individual provider, he/she needs to sign 6 licenses in 6 different judicial systems and will probably have to pay at least 6 different lawyers plus his own!

ELRA considers the production and distribution of these licenses as one of its contributions to the development of LR brokerage, so the licenses are available on the Web (as copyrighted documents) and we encourage all actors to use them. One can get electronic copies from the ELRA Web site.

3.3 Distribution activities and pricing policy

The first two years of activity were devoted to the establishment of the infrastructure and to the identification of valuable resources. This explains the low take off of our sales in 1995-1996. The pricing policy is also a crucial issue that needed careful attention. This had to take into account the fact that we were establishing a new market in which LRs should be traded like any other commodity, bearing in mind the

requirements and restrictions imposed by the provider (or the producer) when it comes to the issue of financial compensation. The ELRA approach is to simplify the price-setting, to clarify possible uses of LRs, and to reduce the restrictions imposed by the producer.

The prerequisite of acting as a broker is that each purchase renders a payment, covering the compensation claimed by the owner of the resource. In general, ELRA is not the owner of the resources, and can therefore only set a fair price in co-operation with the owner. This co-operation in setting the price is often based on conventional pricing methods like production costs. The pricing must also take into account the ELRA distribution policy, which is always to try to offer a discounted price to its members.

In some cases, the providers accept to have their resources distributed for free. This is often the case when production of LRs is already financed by the European Commission or by national governments. When examining the catalogue, you will notice that the ELRA members benefit from price reductions ranging from 10% up to 70% on the public price. Exceptionally, ELRA is able to offer price reductions even without this being financially supported by the providers. This is just one of the services offered to our members, proving that ELRA is unique in its way of offering services and distributing LRs. The restrictions on the distribution, sometimes imposed by the providers, are more often of two kinds: it is either a restriction on the user profile or a restriction on the usage. The providers may limit the distribution to members only or to Europeans only, or they may restrict the use of their resource to research at large or even to academic research. When the restrictions are connected with the type of use, the reason is often that the providers do not want their resource to be used in technical (commercial) development.

The following tables[1] show the situation of the LR distribution via ELRA, for the last four years:

	1996	**1997**	**1998**	**1999**
Members	10	22	156	137
Non members	6	10	24	80
TOTAL	**16**	**33**	**180**	**217**

Table 4. Distribution to members versus non members:

	1996	**1997**	**1998**	**1999**
Research	13	17	90	122
Commercial	3	16	90	95
TOTAL	**16**	**33**	**180**	**217**

Table 5. Distribution with respect to the use of LRs:

	1996	**1997**	**1998**	**1999**
Speech & related resources	13	26	118	97
Written resources	3	4	59	119
Terminological resources	0	1	2	1
Tools	0	2	1	0
TOTAL	**16**	**33**	**180**	**217**

[1] Figures given in those three tables are extracted according to the date of receipt of orders

Table 6. Distribution with respect to the type of LRs:

Major efforts were devoted to the distribution of language resources which led to a 577% increase in our 1997-98 sales over 1996-97. Sales amounted to 180 items sold in 1997-98, compared with 33 items sold in 1996-97. The 1999 fiscal year showed a rise with the sales of 217 items. Whereas our involvement in research and commercial developments was balanced in 1998, items distributed for R&D in 1999 were more numerous than items sold for commercial use (56% vs 44%). Sales of written resources exceeded speech resource sales for the first time (although revenues generated by speech resources represent over 85.9% of the total). Most of our customers join ELRA before buying the LRs (which is justified by our pricing policy).

3.4 Validation and quality assessment

To build up a reputation for the products it sells, ELRA had to set up a system to enable a specification and quality control document to be issued with every product that it licenses. The definition of the validation methodology required co-operating with projects aiming at the production of guidelines, standards, and specifications (e.g. EAGLES, PAROLE, SPEECHDAT, INTERVAL, etc.). Our involvement in validation and quality assessment has seen the release of validation manuals in the area of speech and written resources; these manuals have been made available on our Web site at: http://www.icp.inpg.fr/ELRA/validat.html.

Following its collaboration with the PAROLE consortium, ELRA/ELDA has been entrusted with the work of validating a number of Language Resources (lexica and written corpora) produced in the framework of the PAROLE project. The validation procedure exploited the validation manuals produced by ELRA and referenced as ***"An analytic framework for the Validation of language corpora", "Towards a standard for the Creation of Lexica"***, and ***"A Draft Manual for the Validation of Lexica, release 1.1".***

In order to carry out such work, ELDA had entrusted independent validation units (subcontractors) with this validation task. The subcontracts were carried out with the support of the Experts on ELRA Panel for Validation of written Language Resources (EPV-WLR).

So far the Italian, Danish and Spanish lexica were validated and a validation report is now available. In principle, the validation reports will be part of the Language Resources documentation that any customer should get when inquiring about the data. The producers did their best to take into consideration the comments, suggestions, and criticisms of the validation centers.

ELRA has stated from the beginning of its activities that producing, describing, assuring and improving the quality of language resources is an important task and a success factor for ELRA. In the start up phase of ELRA, it was foreseen to establish a network of Technical Centers which should handle a quality control task referred to as the validation work. In the last fiscal year of ELRA (1997/98) spoken language resources (SLR) have been the most important source of income. Due to this fact, the board decided to start the establishment of the network of validation units with the foundation of a first technical center, namely a technical center for

spoken language resource validation (VC_SLR). It is expected that in the future, other technical centers will be added, whenever adequate.

The procedure to establish the VC_SLR was handled via an open call, widely disseminated. European institutions willing to act as a VC_SLR for ELRA were asked to send an offer to the Board of ELRA.

The main criteria in the selection of an appropriate institution were announced as to be based on the proposer skills to fulfill the following tasks:

- Extending the methodology for describing the Quality and Content of existing SLR;
- Improving the Quality of Existing spoken Language Resources (existing SLR may have errors which could be removed with reasonable effort. The task of the VC_SLR is to build up a procedure to remove these errors; in particular, a procedure has to be established to handle the errors reported by users of SLR);
- Quality Standards for SLR (the VC_SLR has to play a leading role in establishing quality standards for SLR. For this task the VC_SLR has to cooperate with driving forces involved in the production of SLR);
- Validation of New SLR.

Only the Centre for Speech Processing Expertise (SPEX), The Netherlands, submitted its proposal and after the reviewing process, SPEX was selected. SPEX has a deep knowledge of the validation issues as it acted as the validation center for all SpeechDat projects (SpeechDat-M, SpeechDat-II, SpeechDat-East, SpeechDat-Car, SALA, etc.). SPEX started its work by mid'99.

3.5 Commissioning the production of needed Language Resources & market watch

ELRA has issued a series of calls for proposals to help sponsor the production of new LRs, and/or the packaging or customization of existing ones, that are needed by the Language Engineering Community. The purpose of the last call was to ensure that necessary resources are developed in an acceptable framework (in terms of time and legal conditions) by the HLT players. This call targeted projects with short time scales (projects lasting up to one year) and small funding. ELRA funding is to be seen as effective and useful for producers being both tactical in their aims for the targeted market, which means that they do know all about the needs on the specific market, and strategic with regard to what to produce in order to fulfil these needs. The resources to be selected for funding must be in demand on the market and the resources. Market knowledge and contacts with potential providers allow ELRA to always have reliable and useful information on the demands and needs of the users. From its market monitoring, ELRA identified several key speech and written resources. ELRA then categorized and prioritized this set of resources. Preference lists were derived from our market watch and surveys and may gave us some hints about the orientations of the field. The preference lists of particular interest to the Language Engineering community can be itemized as follows:

- SpeechDat-like database (a language or/and an application area not yet covered within the SpeechDat family – 1000 to 5000 speakers),
- Speech database for embedded systems (basically 16kHz sampling, noisy environment, 500 to 1000 speakers),
- Pronunciation lexica (for speech recognition and speech synthesis, including extent of proper names).
- Dialog corpus,
- Enrichment of existing SLRs within the ELRA catalogue,
- Multilingual speech synthesis database,
- Large monolingual corpora,
- Parallel texts,
- Bi/multilingual computational lexica,
- Multimedia corpus,
- Multimodal corpus.

Proposals for other types of corpora were also considered for funding if the resources are necessary for the development of a class of HLT applications. ELRA received over 29 proposals that were screened by a review committee that consisted of the ELRA Board members, a few appointed external experts, and a European Commission (DGXIII - Human Language Technologies sector) representative. This call for proposals led to 8 projects that are currently being co-funded within the LE4-8335 project. The list of projects are as follows:

- *New corpus of written Business English* (Ruslan Mitkov; University of Wolverhampton)
- *Sets of bilingual LR dictionaries for English and Russian* (Vera Semenova-Fluhr; SCIPER)
- *Crater 2 - Expanding Resources for Terminology Extraction* (Tony McEnery; Lancaster University)
- *Italian Broadcast News Corpus* (Marcello Federico; ITC-IRST)
- *Pronunciation lexicon of British English place-names, surnames and first names* (Marc Fryd; Université de Poitiers)
- *Scientific Corpus of Modern French* (Béatrice Daille and Geoffrey Williams; Université de Nantes)
- *German-French Parallel Corpus of 30 Million words* (Wolfgang Teubert; Institut für deutsche Sprache, University of Mannheim
- *Columbian Spanish SpeechDat-like* (Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya)

The call for proposals, the selection of candidates, and all contract negotiations with selected LR production subprojects have been completed during the first year of LRs-P&P. Details of each project start date and end date, along with copies of all technical annexes, are outlined in LRs-P&P project Deliverable 2-1b "Update on Candidate resources for ELRA funding".

Another call for packaging of Language Resources using a similar procedure as for the ELRA’99 call, was

issued on 25 March 1999, with a modest funding from the French ministry of culture through the Délégation Générale à la Langue Française (DGLF). The selection procedure was identical to the one described above, except that the selection committee was set up in coordination with the French ministry of culture. Three projects, out of eight proposals, were selected. These are:

- *Syntsem: Syntactic and semantic tagging of French* (Jean Véronis, CILSH Lab at the Université de Provence and TALANA lab at the Université Paris VII).
- *Annotating grammatical anaphora in French electronic corpora* (Xerox Research Centre Europe, CRISTAL-GRESEC - Université Stendhal - Grenoble 3).
- *Tagging texts to constitute representative corpora* (B. Habert, LIMSI-CNRS and UMR 8503 - ENS Fontenay/Saint-Cloud).

The technical details are available to our members from the ELDA office.

As mentioned above, regular market surveys are conducted by ELRA to monitor the needs of the players in this field. The last one was carried out through questionnaires about the needs, emailed to over 500 contacts. The summaries are given in the ELRA newsletter when appropriate.

3.6 Surveys

ELRA has conducted various surveys in order to better define its action and supply up-dated information on the FLT sector. One of these surveys aimed at identifying the NLP supply in the French market according to its different uses, which gave a general and efficient framework to understand the existing technological and industrial offer in a user-oriented approach. This study, funded by the French Ministry of Research and Higher Education, resulted in a directory of language engineering tools and resources for French. It was followed by an analysis of language engineering tool maturity within the French market. This analysis also dealt with evaluation, which helps to determine tools' and systems' maturity, and potential technology transfers from Research to industrial use and commercialization.

Another worldwide study has been conducted since 1999. It consists of surveys which have evolved and improved over time, and now provide an excellent barometer for measuring LR users' needs. The whole study provides concrete figures for developing a more reliable and workable business plan for ELRA and ELDA, and to determine investment plans for sponsoring the production of new resources. The survey is detailed in a paper by Jeff Allen & Khalid Choukri in the present proceedings.

3.7 Information dissemination, promotion and awareness

Our contribution to information dissemination activities consisted of the ELRA International Conference on Language Resources and Evaluation – LREC. The first LREC conference and its satellite workshops were first held in Granada from May 25 to June 1 1998. It attracted over 500 attendees from over 38 different countries and all the continents. 325 different organizations were present, among which 210 were academic institutions (Universities and Research centers). The program committee selected about 197 papers. Eight pre-conference workshops and a major post-conference workshop about transatlantic cooperation (called MultiLingual Information Management - MLIM) were also organized. The proceedings of LREC'98 and the satellite workshops are available at the ELRA offices. We are very glad to see this second issue of LREC taking place with over 250 papers and about 10 workshops.

One of the other means to make ELRA more visible consists of our quarterly newsletter, issued in French and English. The Web site is another means. We have noticed an increasing number of visits to our Web site and the site is updated on a regular basis, with new resource descriptions and documents of interest to the language engineering community, such as validation manuals.

We have also set up an electronic bulletin to inform our members of our activities. Such a bulletin, which we try to issue on a monthly basis, aims to keep our members updated on our activities and plans, in addition to the announcement of new resources.

4. ELRA collaborations and partnership

ELRA co-operates with other national and regional organizations which are involved in activities relating to those of ELRA. These include LDC (Linguistic Data Consortium), Cocosda, AUF (association of French speaking academic institutions), and plans to cooperate with other emerging organizations.

Another crucial collaboration foreseen by ELRA is with the European National Programs. As of today over 7 countries have started their own program in Language Engineering. The expected outcome includes Language Resources. ELRA has been appointed as the distribution channel for many of them.

5. Conclusion

This paper elaborates on the recent developments at the European Language Resources Association in establishing an infrastructure for the collection and distribution of LRs. It is of paramount importance that existing resources should be identified and made available. It is crucial that we achieve altogether a better streamlining of efforts in the development of new Language Resources that are of interest to "global" players.

6. Acknowledgements

We would like to acknowledge the support of the following organizations: The European Commission, DGXIII, HLT sector; the French Ministry of Research and Higher Education, the French Ministry of Industry, and the French Ministry of Culture, through the DGLF.