An Analytical Model of Language Resource Sustainability

Khalid Choukri, Victoria Arranz
ELDA/ELRA
55-57, rue Brillat-Savarin, 75013 Paris, France
E-mail: {choukri; arranz}@elda.org

Abstract

This paper elaborates on a sustainability model for Language Resources, both at a descriptive and analytical level. The first part, devoted to the descriptive model, elaborates on the definition of this concept both from a general point of view and from the Human Language Technology and Language Resources perspective. The paper also intends to list an exhaustive number of factors that have an impact on this sustainability. These factors will be clustered into Pillars so as ease understanding as well as the prediction of LR sustainability itself. Rather than simply identifying a set of LRs that have been in use for a while and that one can consider as sustainable, the paper aims at first clarifying and (re)defining the concept of sustainability by also connecting it to other domains. Then it also presents a detailed decomposition of all dimensions of Language Resource features that can contribute and/or have an impact on such sustainability. Such analysis will also help anticipate and forecast sustainability for a LR before taking any decisions concerning design and production.

Keywords: language resources, sustainability, language resource lifecycle

1. Introduction

This paper [1] elaborates on a sustainability model for Language Resources (LRs)[2], both at a descriptive and analytical level. The first part, devoted to the descriptive model, elaborates on the definition of this concept both from a general point of view and from the Human Language Technology (HLT) and Language Resources (LR) perspective. The paper also intends to list an exhaustive number of factors that have an impact on this sustainability. These factors will be clustered into Pillars so as to ease both understanding as well as the prediction of LR sustainability itself.

Such descriptive model intends to supply LR producers, packagers, maintainers, data-centers/distributors, users as well as funding agencies with appropriate tools that should help them design a rational and cost-effective lifecycle of LRs. Such model constitutes the second part of this paper, along with the documentation of its first implementation as a web service. Such implementation has to be taken as a risk-management model in the LR lifecycle, which can be used as a tool to anticipate on factors that may not comply with the sustainability requirements.

Rather than simply identifying a set of LRs that have been in use for a while (surviving time passing by and location changes) and that one can consider as sustainable, this paper aims at first clarifying and (re)defining the concept of sustainability, by also connecting it to other domains. The paper will also elaborate a detailed decomposition of all dimensions of Language Resource features that can contribute and/or have an impact on such sustainability.

Such analysis should help us draw a descriptive model of sustainability, usable to anticipate and forecast sustainability of a LR prior to a decision on its development and production (or repackaging).

The creation and use of such resources span several language technology related areas, such as information retrieval, machine translation, speech processing, multimodality applications, etc. In addition to the technological areas, Language Resources are also perceived as a very sensitive issue, touching the sphere of linguistic and cultural identity, with economic, societal and political implications.

Last but not least, this work follows the analysis of the expertise and know-how acquired by the major HLT players within the last 20 years as well as the factual analysis of the lifecycle of LRs that have now more than 15 to 20 years and which are widely known (but not necessarily used) within the HLT community. Some of these expertise and experience were already shared through a number of workshops organized as satellite events to LREC 2008[3] and LREC 2010[4].

2. Sustainability, Self-sustainability: Concepts and Definitions

Sustainability, as widely understood in our field, is the ability of a given resource to survive over time without any explicit and external financial support, which could be referred to as a self-sustained resource. When mentioning such an expression, no reference is made to the availability and use by the wider community, whether

[1] This paper is based on the work carried out within the EC funded project FLARENET and in particular on the work reported in deliverable D2.2 by the same authors.

[2] Language Resources of interest herein are the ones used within the HLT community. Most of the resources are on digital media and format but some may predate the digital area (e.g. dictionaries). LRs refer to data sets that may or may not include "language" components (e.g. images, video streams, signs, sign languages, etc.).

[3] http://www.lrec-conf.org/lrec2008/IMG/ws/programme/W17.pdf

[4] http://workshops.elda.org/lrslm2010/

R&D or industry, no reference is made to its rights management (e.g. licensing), to the ability to be customized to suit existing or new needs, to its "openness" so new users could reshape it and convert it to an operable resource in their environment, no reference is made to its updates, corrections, improvements, repackaging, etc.

According to the Wikipedia and other dictionaries, **"Sustainability** is the capacity to endure. In ecology the word describes how biological systems remain diverse and productive over time. For humans it is the potential for long-term maintenance of well-being, which in turn depends on the well-being of the natural world and the responsible use of natural resources […][5]."

This paper will highlight some of these definitions through examples of "sustained" resources. Many of these "sustained" resources are strongly related to the NLP areas. For instance, many HLT practitioners are familiar with the Bible corpus (and the University of Maryland Parallel Corpus Project: The Bible[6]) that has been prepared and annotated by Philip Resnik and his team. It has been used for a while as multilingual corpora for linguistic research (when other major resources such as JRC-Acquis[7] & Canadian Hansard[8] were not available yet). These Bible corpora used to be freely available and downloadable but nowadays, it is no longer possible (the resources are still available with the 1999 Corpus Encoding Standard[9] as defined by the EAGLES Project[10]).

The distribution of a LR means that it is very likely used (some may be archived for later use but never get effectively used). We can then guess that it is sustainable if it remains in demand. Some resources can be acquired for free and hence there is no guarantee that they are actually used but the possibility to acquire them remains a major indicator on their sustainability.

We have mentioned resources that have been in use for a while, others that have existed also for some time but have not been necessarily used, others that have been developed for specific purposes and that are still available but obsolete (either for some technical reasons such as encoding or the LR has been superseded by a new and better resource). Many resources continue to be maintained, upgraded but many have been packaged as a one-time shot operation. Many have been developed by groups that dropped their activities in the area and no one can trace either the data or the expertise which are no longer supported. Last but not least, we also have those resources which, produced under certain financing programs, are no longer accessible as their developers did not have the financial means to continue any work and thus, they were merely either lost or forgotten in some laboratory corner.

Some resources were important for commercial use and generated revenues that allowed owners to maintain them (but not necessarily).

Some resource developers may still be very active and collect the reported bugs and imperfections. In some cases, these are taken care of but they may as well just be listed without any further action.

This paper aims at elaborating on the "Sustained LRs" versus "Sustained Use", based on various scenarios from which we see that the preservation of some resources can be achieved on a safe and permanent basis, but their usability and effective use cannot be guaranteed. It is hard to tag such resources as sustainable. We will also discuss differences between self-supported and sustainable LRs.

Sometimes, a LR may become *self-supported*. A number of resources can be distributed for a fee and therefore generate enough revenues to allow its owner some regular update, correction, extension, etc. assuming such resource is appropriate for these operations. In some cases, the resource is not commercially valuable but is among the most important resources for R&D and grants a good reputation to its producer. As such, the producing institution could use its know-how (but also that LR) as an argument to join new projects that would increase its R&D manpower.

3. Factors and Features to Consider in the LR Lifecycle

In order to identify the sustainability factors one should first draw a clear picture of the lifecycle of a LR, from the specification phase to the dissemination/distribution phase through the whole production/packaging phase. These three main phases and the lifecycle management could be also applied to repackaging and recycling existing resources to rescue and save valuable pieces and bring them into the new digital/Internet era. In our context most of the resources already exist, mostly on digital media and format, and many of the analysis approach described below would accurately apply.

3.1 LR Lifecycle

The **Pre-Production** phase can be decomposed into various sub-phases that have to do with:

- Production and management of LR documentation (specification documents, reference documents, standards and best practices): these are crucial during a project's specification phase to ensure a long-term use.
- Quality assessment: this allows measuring the compatibility and adequacy of the produced resource with the specified one and it comprises a validation plan and assessment mechanisms.
- Rights, ethics, privacy, consent, and other sensitive issues: all legal aspects need to be cleared out. Unfortunately, this is often neglected and hinders the use of the data by the

[5] http://en.wikipedia.org/wiki/Sustainability
[6] http://www.umiacs.umd.edu/~resnik/parallel/bible.html
[7] http://langtech.jrc.it/JRC-Acquis.html
[8] http://cw.routledge.com/textbooks/0415286239/resources/corpa3.htm#_Toc92298948, http://www.isi.edu/natural-language/download/hansard/.
[9] http://www.cs.vassar.edu/CES/
[10] http://www.ilc.cnr.it/EAGLES/home.html

producer, and, at a later stage, its distribution and use by third parties. It is crucial that the project reviews all background knowledge used within the project and the related copyrights, patents, other background ownership, ethics, sensitive issues like consents, etc.

Production is the most expensive phase and requires very good planning and attention, particularly if the LR is produced *ex nihilo*. The first stage is to ensure that the final data will comply with the specifications. We will decompose this phase into various steps going from Information Dissemination along with drafting and agreeing the specifications, production procedure, quality assessment plans, description of intermediate progress, and possibly some prototyping of potential applications, etc.

Once the data is produced, packaged and extensively used by the researchers that commissioned it, they should be encouraged to share their results with the community. This would encourage others to use it. By publishing their research papers, users will also make the LR more known. Since the last LREC (May 2010), it is also essential for LRs to be part of the new instrument set up by ELRA and FlaReNet, the LRE-Map[11].

Moreover, one should ensure that the data is encoded following the right encoding practice (for most languages, one should use Unicode as the character encoding standard, available for almost all languages). This is also applicable to file format, mark-up language, storage and packaging.

In addition to the use of such data and file standards, one should provide adequate tools for the human users to view, read, listen, visualize, and manipulate the data.

The next phase is the **Post-Production** phase. The main task herein does not relate to the "internal" exploitation by producers/owners but rather to the wide dissemination for sharing with the whole community. In order to start this phase, it is crucial to be sure that the producer/owner did check all legal aspects[12] in advance and that he/she knows which distribution strategy and policy to implement.

Another very critical aspect to address by data owners is the data identification by the potential users. The most valuable resource is useless if no one can discover it. Such identification implies that the information on the LR is compiled and widely disseminated (even beyond the usual HLT community). The compilation of information also assumes that the data is well described through the use of adequate metadata and that potential users have an easy and efficient access to the documentation but also enough samples to assess their usability and relevance in a particular context.

In order for the users to identify a given LR in a persistent way, it is also mandatory that the resource bears a Unique and Persistent Identifier. ELRA is working towards a consensus within major HLT organizations to attribute an identifier to LRs similar to what is done with other digital objects [13] . Such unique identifier is the ISLRN (International Standard Language Resource Number), whose description and presentation can be seen in the current conference too (Choukri et al., 2012).

Very often, owners assume that a LR can be identified by a web URL (Uniform Resource Locator*)* where the data is either documented or even stored. URLs change often according to the hosting institution policy and infrastructure. A unique identifier, independent from the storage place may improve the LR "identifiability".

Identifiability is a crucial factor that weighs a lot in the sustainability score of LRs. It requires that:

- Information is compiled and disseminated on the LR;
- Scientific and technical publications are encouraged with accurate reference to the LRs, on the major conferences and journals and on the LRE-Map;
- Accurate and common metadata sets are used;
- A Unique and Persistent identifier is assigned to the LR (or requested from some data centers);
- Metadata harvesting is allowed even if trustable and reliable cataloguing is preferred.

It is of utmost importance for the use of the LR by third parties, that they have to access its content so they can incorporate it in their own environment. We assume that most of the users expect to have access to the content in an "open" mode. No one expects to get an encrypted file with API but rather "plain" data that one can manipulate freely (given the terms of the agreed-upon license). A new trend is emerging in which some resources are made available as Web-services.

Such scenario allows users to access some resources stored on a remote server and use them through some specific APIs. This approach is often used for testing purposes (before acquisition of the database) but also to ensure that users are exploiting the latest version of the resource. As one imagine, this creates a dependency on the LR owner infrastructure (its servers and web-services) and may be incompatible with the user strategies (in particular for resources that have to be exploited without Internet access such as new PDA applications).

In addition to some online storage, the LR could be safely stored on some physical devices or media such as CD-Rom, DVD, hard disk, USB key, etc. It is important to bear in mind that most of these devices have a limited lifetime (their longevity is from a few years to a decade). It is therefore crucial to migrate the full data package regularly to new media.

In addition to this, the owner (or the data manager) should ensure a serious backup plan that guarantees that destruction of a copy or corruption of the content does not imply the loss of the LR. Usual process applied to valuable and expensive data (including financial ones) should be applied. In particular multiple copies and off-site location should be adopted for storage. Automation of such process and its regularity may impact the preservation and thus sustainability.

It is critical to store safely the "Raw" (primary data) that is

[11] http://www.resourcebook.eu/LreMap/

[12] For simplification purposes, we use the term legal also to include ethics, privacy, sensitivity, and other related issues.

[13] For more details, see http://www.doi.org/

the basic collected information (raw recordings, html text pages, etc.) and migrate them forward in time on a regular basis, in addition to any processed data (analyzed speech, transcribed speech, cleaned texts, annotated texts, etc.). Redoing part of the work, e.g. transcriptions of 2,000 speakers, represents a substantial effort but not as important as recording 2,000 speakers all over again.

It is therefore essential to consider the Sustainability of Preservation as a key element in the sustainability plan.

It is clear that many LRs are also of high commercial value and require substantial development efforts. One may consider that they should be treated like any other commodity following rational market rules and therefore considering pricing them at a level indicated by market demands.

The market today (2010) has a clear and strong bear trend. In particular most of the academic users expect LRs to be free of charge. A high price will definitely limit the use and dissemination of the LR while a free of charge/low-fee LR, if it does not guarantee such usability, may probably help it.

It is important to document the LR with respect to the project and framework in which it is produced and in which it has been used. Reference to such projects and areas of use have some impact as users may be inspired by analogies with resources produced in the same or similar projects.

It is also crucial to document the languages, topics, applications, projects, for which the resource has been initially designed and has been used. Using such resources could be boosted by some teams' publications of the performances they achieved on particular systems.

The language addressed by the resource has also an impact on the surviving of the LR and its usability. If the language is part of a mainstream then its chances to survive and develop are important. Basically the ones under spotlights today are either those for which lucrative applications can be deployed, those for which the size of potential market is impressive and investment are required for tomorrow's applications, and those for which geo-strategic considerations require heavy investment of particular agencies. Examples could be respectively English/Japanese, Mandarin/Hindi, Pashto/Urdu. In some cases, national agencies understood that funding Language Resources was a prerequisite to allow their culture to survive and thus devoted some efforts to that (e.g. Basque and Catalan), allowing the development of highly attractive technologies even for these less-lucrative and less-resourced languages.

The areas in which such LR is usable is also of paramount importance to the surviving of LR. If it is used in some mainstream areas (e.g. today these are MT, IR, Speech transcriptions), this will give it a wider audience and potentially more users.

Some resources may end up being a standard like Aurora, MLCC, MULTEXT (ELRA), TI digits, TIMIT Acoustic-Phonetic Continuous Speech Corpus, UN Parallel Text (LDC), CLEF and NIST/TREC evaluation packages and used by many PhD students in their work, extending the LR life.

3.2 Life of the LR on the Long Term

3.2.1. Role of Data Centers and Archiving Houses

An archiving house of data and/or of metadata may play an essential role in the preservation and promotion of a LR. The role of data centers and archiving houses are of different natures. A data center can simply archive a LR and thus play a role of an off-site backup center (an important though a passive role). The data center can also play a role of a distribution and promotion center that would ensure that the data is promoted within the right communities, made available through adequate means and appropriate licenses (including a copyfree and no-licence option). In order to play such role in a reliable manner, centers have to show their experience and expertise for these tasks but also have some credibility and longevity in long-term preservation and access. In principle, the data centers have also to address issues like assignment and management of the Unique Persistent Identifier of the LR, of the LR versioning, migration of resources to new devices and new format whenever current hardware may become unsupported. They have to consider all issues related to data backups, off-site backups, regularly migrating LRs to new infrastructures (new servers and web tools, new search engines, new access/delivery modes (e.g. CD/DVD, new hard-disks, ADSL to Optical Fiber). In order for such a data center to comply with these requirements and to fulfill these duties, it has to be sustainable itself! This means, within today's landscape that either it enjoys a long-term institutional support, including a financial one (in some case this is the case of public institutions), or enjoys a serious record of community support and backup (case of associations and other institutions, e.g. ELRA, LDC), or finally enjoys a profitable financial situation through revenues generated by the tasks mentioned above when applied to the archiving of such LRs.

In the first case, the public institution may have its own roadmap, dictated by its governing body. The association strategy and policy is, in principle, dictated by its members who are often owners/providers/users of LRs. The company has its own stockholders that expect it to generate profits. The shareholders may also set a different policy direction to ensure that the company performances are consistent with their priorities. The debate on efficiency and cost-effectiveness of public institutions, semi-private ones, and corporations transcend this report but all the arguments can be brought up herein. The major question is about how to fund the operations required by the tasks of a datacenter: through public funds (in which case LRs do not have to be self-supportive) or through private ones (in which case some LRs have to generate enough revenues to sustain the non-lucrative resources).

There is also a debate about the capacity (and the strategy) of data centers to implement archives and catalogues that are "ready" to be harvested by other centers.

If the data center also takes in charge the "publication" of the resource (formatting, validation, packaging, …) then it should adhere to best practices for the documentation (e.g. Metadata) and the data (format, encoding, media storage).

New trends of established archiving houses for metadata

are well represented by examples e.g. OLAC [14] and META-NET[15].
OLAC (Open Language Archives Community) is about the creation by many international partners of a worldwide virtual library of language resources through "the development of consensus on best current practice for the digital archiving of language resources, and developing a network of interoperating repositories and services for housing and accessing such resources".
META-NET is an EC FP7 Network of Excellence dedicated to building the technological foundations of a multilingual European information society and an important part of the project is about the setting-up and running of an open LR infrastructure to archive, use, share, exchange, etc. LRs.
Data centers are more viable and "sustainable" when they offer packaged repository solutions based on open standards that enable providers to set up their own repositories if they do not wish to join the core repository and catalogue. By doing so, such institutions would both compete and collaborate.
These initiatives highlight the new trends (open, distributed, etc.) and may impact the sustainability factors over the new decade.
In all cases, the owner/producer of the LR should retain its copyright and ensure that he/she can move the data elsewhere if that data center fails to play its role efficiently.
Another recent major action, initiated in the framework of the FlareNet project, and in cooperation with ELRA, is the LRE-Map as introduced in LREC'2010. The LRE-Map aims at collecting information (metadata) about the language resources in conjunction with scientific and technical publications that elaborate on various issues related to the design and specification of the resources, its quality (validation assessment), its use within particular projects and topics, its use to evaluate existing technologies, etc. Such LRE-Map (already available at LREC'2010 and COLING'2010) should improve the information dissemination of LRs and somehow have an impact on their sustainability.

3.2.2. Maintenance and Support over time

As stated above, many resources continue to be maintained, improved, corrected and/or upgraded by the owners or by the community when the resource is made public through some licensing schemas. For instance, if the licenses grant users more rights than just "redistribution" of unchanged LR and in whole (e.g. the rights to modify, remix and build upon the LR, in principle as long as they credit the owner of the original resource), one expects a community to be established (as we see with the open source software communities) with forums, reporting boards for bugs and errors, recommendations, etc. to take care of the resource.
Like most of the open source development, if the LR is not enjoying strong community back-up, its support may simply and quickly vanish. On the contrary, one can see a large number of releases but also different and diverging versions if the LR succeeds in federating a large number of users.
The example of WordNet is very significant. A strong community support helped making this resource available and widely used. Such community support is hard to predict and hard to model. A crucial resource for a language could be considered as very critical and can boost a community support (e.g. WordNet, American National Corpus) that would adhere to common policies and practices (this is somehow the spirit in the software open source community). One can imagine a sustainable community support that has an impact on the sustainability of the LR.
On the contrary, some resources may as well emerge in a framework of competition between two strong teams, each advocating holding the "truth". This may lead to different (but very similar or close) resources. For instance, there are two resources for French[16] WordNet[17] .
How can one ensure such community support over time?
In the case of a resource that is not shared under the above principle but rather protected by its owner, we notice that many such resources have been packaged as a "one-time shot operation". One release of the data is made available and no improvement, update, upgrade is foreseen. On the contrary, some resource developers are very active and collect the reported bugs and imperfections. In some cases these are taken care of but they may as well just be listed (very often as a courtesy to the users but not as an institution commitment). Some resources have been developed by groups that dropped such activities and no one can trace either the data or the expertise so these are no longer supported. Some resources are important for commercial use and generated revenues that allow owners to maintain them (but not necessarily).

4. Other Factors

In the literature, we also encounter terms that express slightly different concepts regarding the factors that have an impact on LR sustainability. Some that we have not directly listed above are discussed herein.
Some experts argue that "scalability" is an important factor. In most of our resources for Human Language Technologies, we assume that a resource has a right size when a first version is released and many will have no extension. We also assume that versioning would address issues like correction of bugs but also updates (including in terms of size/scale) even if this is not necessarily consistent with "scalability" capacity of a given resource.
"Interoperability" [18] is another term that we tackled through a number of factors such as adherence to best-practices and standards, Quality assessment and Quality validation report, LR Format, Encoding, Content, LR Portability across languages, environments and domains, metadata, etc.
Another expression encountered is "Viability". This has been expounded over items like usability assessment,

[14] http://www.language-archives.org/
[15] http://www.meta-net.eu/
[16] http://alpage.inria.fr/~sagot/wolf-en.html
[17] http://catalog.elra.info/product_info.php?products_id=550
[18] Interoperability: capacity of a LR to be usable by different systems, in different environments, capacity to exchange data with other resources. In general, it refers to the ability of two or more systems or components to exchange information and to use the information that has been exchanged.

accessibility, preservation of media, etc.
"Equitable" is also a keyword encountered in the UN reports. In our area, this has an implicit meaning when we refer to the less-resourced languages. It is fair and "equitable" to allow access to the data for researchers from these language communities under a very specific licensing schema, in particular when it comes to the "commercial" part of it. It is also "equitable" to share part of the revenues, if any, with the local community that helped with such development. This is highly contradicted today with the novel approaches used to collect data such as the Amazon Mechanical Turk. This has to be treated even more sensitively for indigenous languages that have no local R&D task forces.

5. Intrinsic versus Extrinsic Sustainability Factors

Different resources follow lifecycles of different natures and many of these steps may require, for their preservation and long-term use, different kinds of organizational set-ups and may partly involve different stakeholders. Our sustainability model will be based on such general assumptions, although we make a clear distinction between (efficient) management of LR lifecycle and sustainability.
Some factors that have an impact on the life of a LR are either solely related to the LR itself or related to external contexts. The first ones are explicitly inherent and underlying what the dataset is, its nature and the content that is built in. The others are extraneous factors that relate to the general environment of the resource but that still have an impact. We will refer to these are Intrinsic versus Extrinsic factors.
A typical intrinsic factor would be the nature of the Language Resource and the type of language knowledge it represents. For instance, Wordnet is considered as a major resource for NLP and has been in use for decades (for the Princeton version) and over 60 languages have emerged since then, particularly after the success of the European project EuroWordNet.
WordNet is a "large lexical database of English in which nouns, verbs, adjectives and adverbs are grouped into synsets (sets of cognitive synonyms), each expressing a distinct concept". WordNet has been freely and publicly available for download. The WordNet structure and content raised very controversial debates within the community and some contradicting initiatives were launched but did not encounter the success of WordNet. Such longevity is due to the nature of the data represented (a semantic lexical database), to the original language (US English), to its free (free of charge) availability and the easy license that governs its use (a license that grants permission to "use, copy, modify and distribute [this software and database and its documentation] for any purpose and without fee or royalty").
Five factors are implicitly listed herein: the importance of the NLP area (Lexical/semantics), the database format (synsets), language (US English), the availability-for-free and the license. Some are intrinsically related to the database (area, format, and language), others are extrinsic and could be modified and changed without touching the resource itself and at anytime (fees, license).
A typical intrinsic factor is the language represented by the resource. Within the WordNet family and given the example of the English WordNet of Princeton, it is hard to predict how many copies will be distributed for a language like Korean or Russian but also how long such resources will remain in use.
We have seen that a resource for languages like Arabic or Basque would have more chances to be preserved, used and re-used (even though it may not be updated or improved) than a similar resource for a language like German or Norwegian. We have a few examples of corpora that illustrate accurately such statement. We can draw a list of languages that are either:

- Very lucrative (and then most commercial and large R&D centers will be highly interested e.g. English, French, Spanish, Portuguese, Chinese, etc.).
- Geopolitically sensitive (and then a large number of research groups focus on them raising the interest of major stakeholders like data centers and archiving houses, e.g. Arabic, Urdu, Pashto, etc.)
- Part of a minority language group (Basque, Gaelic, etc.) with either many specificities that makes it attractive to researchers, or politically sensitive and then strongly backed up by a community wider than just NLP.

A license is a good example of an extrinsic factor. In other reports and documents we have elaborated on the variety of licensing schemas that allow obtaining and using a given LR. Such licenses should be customizable and also adaptable over time to suit better the needs of the users and the requirements or expectations of the right holders[19] that may evolve over time. The initial license could be very restrictive to allow the owner to derive as much benefit as possible (financial, scientific, etc.) but after a time period may become more permissive[20] and tolerant to allow for a wider use. For a number of resources, the use of Creative Commons Licenses is highly recommended unless one can simply let the related work fall into the public domain by adopting a copyleft approach.
The "pricing" (or distribution fees) is also another extrinsic factor and ELRA has been working on this since its foundation (reference: internal pricing rules by ELRA). The LRs bear two specific features of (often) contradictory natures: they are part of our cultural heritage and all technologies related to that give access to information, culture, etc, and some may argue that LRs should be treated in a particular manner. The other feature would consider the LR as any other commodity and thus could be of some commercial value to which owners should apply a rational pricing policy. It is important for the owner to understand and define such policy in principle according to market rules. One can envisage a price based on production costs (e.g. one needs to sell N copies in order to recoup the invested funds, which is very often hard to assess for resources produced over decades in varying academic environments) or review the market

[19] LREC Workshop on legal issues (http://workshops.elda.org/lislr2010/)
[20] Not to be confused with the "permissive free software licence": copies and derivatives of the source code created under permissive licenses may be made available on terms that are more restrictive than those of the original license.

demand and guess how many potential users are prepared to pay for it. Since mid-2000, the market showed a clear and strong bear trend in addition to the wide spread of Internet that boosts the feeling that such resources should be free of charge. A high price will definitely hinder wide use and dissemination while free-of-charge/small-fee will not guarantee such usability although it could boost it.
These are two concepts (rights/licenses; price/fee) that are extrinsic and that could be adjusted in time according to their environment evolution without touching the resource itself.

6. Chart of LR Sustainability Factors and Means

Let us summarize the concepts introduced above and also cluster some of the factors into categories for which we could elaborate a sustainability scenario. The major factors we highlighted above are summarized herein in 20 issues/factors that we feel have an impact on sustainability:

1. LR specifications (incl. references to best-practices & 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
7. LR Portability across languages, environments and domains
8. LR packaging (compilation of all pieces together incl. resource, documentation)
9. Rights to be granted and Licenses
10. Data identification, metadata and LR discovery
11. Versioning and referencing of the LR
12. Usability assessment and Relevance
13. Accessibility of the LR (LR package, medium)
14. Accessibility of LRs in an "open" mode
15. Preservation of the 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 & areas
19. Maintenance and support over time
20. Role and Impact of data centers and archiving houses

Some of these items are direct impacting factors. Others are important means, facilitators and/or actions that have a strong impact on some factors. These have been categorized and clustered into major/minor factors but also labeled as intrinsic or extrinsic. Different weights have been assigned (ranging from 1 for "minor" and 5 for "major" factors) so as to know what their impact is. For the full computation, please refer to (Choukri & Arranz, 2010).

7. What are our own Sustainability Pillars?s

It is crucial to understand better and also assess the weights and relative importance of all these factors and facilitators on LR sustainability.

In order to determine the most important ones, let us define "Pillars" that consist of clusters of the main dimensions and which need to be reconciled. Examples are depicted in Figure 1.

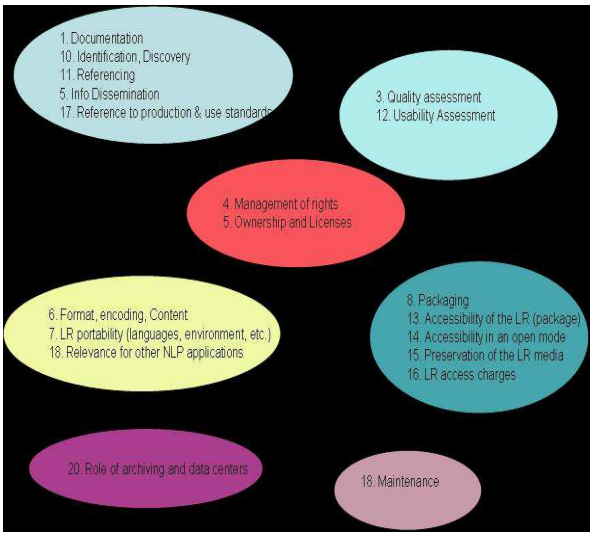

Figure 1: Clustering of the various factors and means into groups (Pillars)

This view has been expressed as an illustration using overlapping ellipses indicating that the pillars of sustainability are not mutually exclusive and can be mutually reinforcing. We will also see that a cluster of pillars will also induce some sustainability of access, sustainability of preservation, sustainability of community support, etc.
The overlapping is hard to illustrate (see Figure 2) but one can imagine how these dimensions are to be combined in order to enhance the sustainability of a LR.

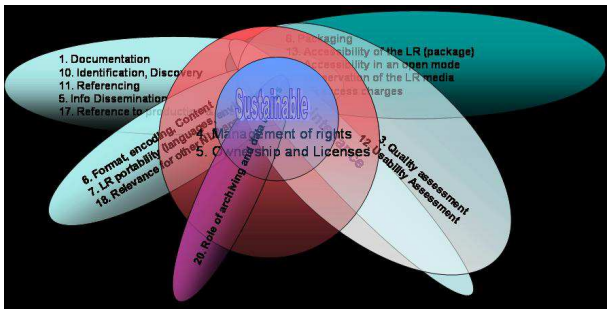

Figure 2: Combination of the various Pillars to define the sustainability area

From the diagram given above, we can draw a clear picture on which core factors and/or means have an impact on the sustainability and their related topic. We can distinguish a LR model sustainability, a sustainability of access, a technological sustainability, a preservation sustainability, etc. As we can see, these items are not necessarily exclusive. It is rather the combination of several of these impacting factors and means that leads to a global sustainability.
The LR sustainability model requires that the factors that have an impact on the LR production specification, on its documentation, on its referencing and promotion, etc. are

treated on the most efficient and appropriate way. For example, the production specifications should adhere to some best practices as adopted by major players, big projects, and international initiatives, to enhance its usability and hence its sustainability chances.
If we look at the sustainability of access, the major factors are those related to the access to the LR package, access in an "open" mode, access to a preserved media, access charges but also to the factors related to the legal issues (management of background rights, copyright and ownership, and licensing). It is also important to have access to the documentation, to some maintenance and support, etc.

8. Sustainability Model, a Risk Management-like Model

In order for these factors and means to be accurate parameters for the assessment of the sustainability of a LR, we need to turn this into an analytical model. For instance, some factors are independent, some means depend also on other factors and means, etc. Our assumption is that in modeling these issues, we consider a risk management model that is based on "educated guesses", intuitions, as well as some statistics about resources that have been in operation for the last 10-20 years. The basic role of the model would be to help production players assess the sustainability of the LR to be produced, and optimize all factors and instruments that have an impact on this sustainability. Although such a model would help foster choices among several possible alternatives and reduce subjective approaches, the only way to evaluate the sustainability of the LR will be *a posteriori* and would be based solely upon the results of the analysis of the LR life.
From our experience we can draw some conclusions about the important factors and assign them some weights and thresholds. A global score could be computed with respect to a sustainability value.
From the main table of sustainability factors, one can see the important ones and the minor ones. We can also assume that some of these factors, in addition to their weight with respect to sustainability, have a threshold value that would seriously hinder the LR sustainability.
If, for instance, a LR has a serious drawback in its sustainability cluster consisting of [1. Documentation, 10. Identification, Discovery, 11. Versioning & Referencing, 5. Info Dissemination, 17. Reference to production & use standards], then this would certainly jeopardize its sustainability.
On the other hand, if the cluster consisting of [8. Packaging, 13. Accessibility of the LR (package), 14. Accessibility in an open mode, 15. Preservation of the LR media, 16. LR access charges] is well done and the resource is backed up by a well-known institution then its sustainability could be higher.
The idea of this model would be to act as a simulation tool of the sustainability probability for a given LR.

9. Conclusions

This paper aims at providing a detailed view of "sustainability" in LRs through the analysis of LR lifecycle. Throughout the expertise obtained these past 20 years, we analyse the factors involved in the production of long-term and sustainable LRs while defining a descriptive model. This model should allow LR producers, maintainers, users, etc. design a rational and cost-effective LR lifecycle.
An exhaustive list of factors and instruments that have an impact on sustainability of LRs has been drawn. We have defined a first framework to model the sustainability of Language Resources. Some examples have been illustrated with existing resources for which we can judge on sustainability. It is clear that the new trend is for a public or public-private partnership for the production of LRs but also for open, freely available ones. This new trend will require a coherent action supported by appropriate public policies. It is crucial that such policies shall endorse "sustainable management" of LRs through experienced centers but also fully integrate sustainability concerns into their decision making and management practices.
We have seen in our model that LRs can be sustainable from various perspectives: sustainability of access, sustainability of preservation, sustainability of use. A large number of players can play a role herein and coordination may become very crucial. The involvement of all stakeholders is required at all LR life stages. It is crucial, for projects like FlareNet but also for well-established organizations like ELRA to drive the attention of the whole community towards the challenges of LR sustainability.
An important external factor that has an impact on this issue is the strengthening and consolidation of the HLT players. Strong HLT players will more easily consider the LRs as their sensitive assets and implement the right policy for sustainability. This, of course, requires the sponsors ability (in addition to HLT major consumers) to sustain actions over time.
A holistic and integrated approach to sustainability should be taken into account in the LR planning and development, involving all stakeholders.
Last but not least, sustainability is not a "frozen" quality, acquired once for all. It is important for the key players (but mostly data centers) to undertake a continuous monitoring (somehow similar to the Universal Catalogue of ELRA and the new LRE-Map) of all the factors listed herein. Some extrinsic factors may need adjustments and carefully monitoring them would alert the right player about when changes are to be made.

10. Acknowledgements

This work has been financed by the EC through the Grant Agreement No. ECP-2007-LANG-617001.

11. References

Choukri, K., Arranz, V (2010). Identification of mature self-sustainable LRs vs. areas to be sustained & sustainability factors. FlaReNet Report.

Choukri, K., Arranz, V., Hamon, O., Park, J. (2012). Practical and Technical Aspects for Using the International Standard Language Resource Number. Proceedings of the Eighth International Conference on Language Resources and Evaluation. Istanbul, Turkey, 21-27 May 2012.