ECP-2007-LANG-617001

FLaReNet

Deliverable 8.2a
Blueprint of Actions and Infrastructures

Deliverable number/name	*D8.2a – Blueprint of Actions and Infrastructures*
Dissemination level	*Public*
Delivery date	*31 August 2009*
Status	*Final*
Author(s)	*Nicoletta Calzolari, Claudia Soria, Núria Bel, Gerhard Budin, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Antonio Toral*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Table of Contents

TABLE OF CONTENTS **2**

1 INTRODUCTION **3**

2 ASSESSMENT OF CURRENT SITUATION **4**

3 RECOMMENDATIONS FOR ACTIONS IN THE HLT FIELD **6**

3.1 HLT STAKEHOLDERS 6
3.2 FUNDING AGENCIES AND POLICY MAKERS 8

4 RECOMMENDATIONS FOR INFRASTRUCTURES IN THE HLT FIELD **10**

4.1 THE MAIN MESSAGE: AN OPEN RESOURCE INFRASTRUCTURE TO BOOST THE LANGUAGE TECHNOLOGY FIELD 10
4.2 HLT STAKEHOLDERS 15
4.3 FUNDING AGENCIES AND POLICY MAKERS 16

1 Introduction

This document contains the first *Blueprint of Actions and Infrastructures*, as developed by FLaReNet after one year of activity. It outlines the first contribution to a comprehensive framework or plan of action for the future. The meaning of comprehensive here is twofold. In the first sense, by comprehensive we mean that the addressees of this document belong to a large set of players and stakeholders in Language Resources and Technologies, from individuals to research and education institutions, to policy-makers, funding agencies, SMEs and large companies, service and media providers. In its second sense, we also intend that the Blueprint series is intended to cover a broad range of topics and activities, spanning over production and use of language resources, licensing, maintenance and preservation issues, infrastructuresfor information and resource sharing, evaluation and validation, and interoperability.

This document is the first in a series, each coming out every year, and as such its purpose is to provide the addressees with a first concise and clear set of recommendations concerning actions to be undertaken for the advancement of the field of Language Resources and Technologies (LR&T).

To this end, this document aims to serve as a tool (or "blueprint") to support organizations, institutions, funding agencies, companies, and individuals in planning for and addressing the needs of the LR&Ts of the future.

Recognizing that the development of the sector of LR&Ts is influenced by various factors, these recommendations suggest an approach in which all interested stakeholders come together to forge partnerships to promote LR&T. These recommendations can be taken into account by any player, whether on a European, National, local, or private level, wishing to draft a program of activities for his/her own communities.

Together, and under the umbrella of a shared view of actual priorities, a future can be shaped in which a common market for Language Resources and Technologies is created through coordination of programs, actions and activities. While there has been considerable progress in the last decade, there remains a significant challenge to overcome current fragmentation and imbalance inside the LR&T community. We consider the process for developing the *Blueprint* as important as the resulting document itself; it is through the act of engaging the community in discussing, challenging, justifying, and reconciling their individual and collective views, experience, and concerns that the *Blueprint* has come to fruition.

As such, the development of the series of FLaReNet *Blueprints* will be the result of a permanent consultation that FLaReNet has opened inside its community and outside it, through connections with neighbouring projects, associations, initiatives, funding agencies and government institutions.

The process includes input from more than 240 people and 80 institutions from 31 different countries.

As a result of this process, these *Blueprints* are considered to reflect a consensus on the active living interests, needs, concerns, and values of a wide spectrum of players in LR&Ts.

More in detail, this first *Blueprint of Actions* further elaborates on the FLaReNet Action Plan (D8.1), and prepares the ground for a plan of actions for LR&T through awareness, support, infrastructural and research activities. The document includes key recommendations as they emerged from the FLaReNet community of experts, mostly during the various FLaReNet meetings and events, but also from external events that were endorsed by FLaReNet, or from projects/initiatives/organisations with which FLaReNet has close relationships and contacts (e.g. ELRA, LDC, CLARIN, SILT, CyberLing, LanguageGrid, AFNLP, ISO, COCOSDA, ALTA). Those recommendations were further refined by the members of the Steering Committee.

This first version of the *Blueprint* lays down the basis of a comprehensive plan of actions for the various typologies of actors involved, mainly by listing the recommendations addressed to two broad categories of intended recipients: HLT organisations at large (which include producers and users of LR&Ts) and funding agencies and policy makers.

Dependencies and priorities among recommendations, as well as an indication of timing, risks involved, and in general, a systematic view of the future of the sector, will be specified in subsequent versions of this Blueprint.

According to a well-established policy in FLaReNet, this document will be made open for comments and contributions in collaborative mode on the FLaReNet web site.

2 Assessment of current situation

Work conducted so far in FLaReNet has contributed to draft a first portrait of the current situation in the LR&T sector, in particular for what concerns the types of players and resources (WP2), the various needs for standardisation according to the different communities and the obstructing factors to adoption of standards (WP4), an overview of current practices in evaluation and validation of LR&Ts (WP5), and a review of the innovative methodologies being implemented for the automatic development/processing of LRs (WP6). In addition to the activity of the work packages, input has been collected from a number of events, either organised or co-organised by FLaReNet.

The following is a shortlist of facts that concisely hint at the situation of the LR&T sector as it has emerged from FLaReNet observation.

- Re-use and re-purposing of data is hindered by lack of common data representation
- Documentation of language resources is generally poor
- Clear and easy-to-reach information about resources and related technologies is lacking
- There are too few initiatives around the BLARK concept for European languages
- Little concern is given to the issue of data preservation
- The legal framework is far too complex, and in particular:
    - Licence models especially suited to LRs are lacking
    - Legal protection modes are different across Europe
    - There are different strata of intellectual property rights
- Sustainability for linguistic tools and language resources needs to be increased
- LRs need to be mantained, in terms of bug reporting, updates and improvements

- More efforts are needed to solve the problem of how to automate the production of the large quantity of resources required, and at the same time how to ensure the necessary quality to get acceptable results in industrial environments
- The evaluation of automatic techniques for LR production is of variable quality. Comparisons among techniques should also be carried out to better assess each of them and their strengths and weaknesses, fostering a greater development in the research on these fields
- Much of the research on automatic acquisition of LRs has focused on small-scale experiments and therefore their usability in applications is largely yet to be demonstrated
- It is very difficult to find information about the characteristics of the language resources that industrial applications use, as well as about the size and granularity of the information contained
- Standardisation is at the core of interoperability. Standardisation issues currently show substantial convergence of opinion and practice, which needs now to be supported to become operational
- LR standards are:
    - too much oriented towards academic/research purposes, not yet mature enough for industrial applications
    - too difficult to understand
    - too abstract, lack concrete examples for implementation, lack of user scenarios or user guides
    - too isolated, existing only on paper but not integratable in digital workflows,
    - too cumbersome to implement, no return on investment in sight for implementers
- Industry-born standards are:
    - too much driven only by specific needs and lack long-term vision
- Given the breadth of current landscape of LR&Ts, a “cultural” change is needed in the sense that there is the need to find ways to monitor how resources are used, to register the resources used or created, to introduce the notion of “publishing” resources and to get academic credit for resources that have been made available.

3 Recommendations for Actions in the HLT field

3.1 HLT Stakeholders

HLT stakeholders include producers, users and developers of Language Resources and Technologies, both academic and industrial.

3.1.1 Resource production and use

- Provide documentation of the produced resources, covering at least the following aspects (metadata): owner/copyright holder, format and encoding issues of the data and the files, languages(s) covered, domains, intended applications, applications in which the data was used, formal parameters that have to be verified, reliability of any annotation that is included
- For documentation, adherence to practices followed by major data centers is advisable
- Ensure quality of language resources (LRs), for instance by performing a basic quality assessment, to ensure that the minimal critical dimensions are documented: availability/reliability of information on technical, formal issues such as media, number of files, file structure(s), file names etc.
- Annotated resources should be provided with a detailed documentation describing the annotation procedures which have been developed in the annotation process of the LR
- Promote the development of new methodologies for assessing the annotation quality of LRs, in particular for semantic annotation
- Information about whether the resources acquired are actually used or, the other way around, of what are the particular characteristics of the actually used resources, needs to be made public.

3.1.2 Interoperability issues

- It is important that commonly accepted practices (best practices, de-facto standards or standards, when available) are used for the representation and documentation of data
- Not only are data formats to be standardised, but also metadata
- Standards need tools that support them , to promote and ensure their adoption
- LR standards have to be made *more operational* (both, existing ones and those under preparation), with a specific view on different user communities – most users should not or do not want to know that they are using standards, they should operate in the background and they should be "inherent" to the language technology tools or more generic tools they use
- A crucial step towards consistency and interoperability for a global information exchange is the definition of a work environment for *data category definition and management*

- Aim at new forms and manifestations of standards, as *embedded standards*
- For each standard, *return on investment* and possible *motivations* of users should be elaborated together with potential or real users (early adopters)
- Focus in the *short term planning* on those areas where there is enough consensus so that chances are high that a widely accepted standard can be published in a short period of time
- Increase the *acceptance* of LR standards (and the need for them) in different communities, both research and industry communities, and to directly involve user communities in creating standards
- Analyse the needs and requirements for harmonisation of existing standards
- Develop a *strategy* for LR *standards creation*, taking into account aspects such as: bottom-up vs top-down approaches with an interactive process model needed, and modular component standards rather than a single monolithic standard for all of LR
- Standards maintenance should be a process of *change management*, ideally *in real time*
- Inform more pro-actively on best practices in implementing standards and in successful corporate language standards.

3.1.3 Licensing, maintenance and preservation

- Prevent loss of data along the years, by ensuring appropriate means for data archiving and preservation
- Avoid "home-made" licensing models. When drafting a distribution license, carefully think of making it suitable for subsequent re-use and re-distribution of the resource. Adhere to practices used by distribution agencies whenever possible
- Whenever possible, ensure appropriate means for maintenance of LRs.

3.1.4 Evaluation and validation

- Work on common and standard evaluation procedures, taking into account normalization for comparison. Techniques should not only be evaluated on scientific grounds, but also by their impact in real scenarios of NLP applications
- Develop tools for automatic validation (fault detection (clipping, noise...), detection of segmentation errors, of weak annotations, confidence measures of speech transcriptions)
- Investigate different solutions for addressing the problem of task- vs. application-oriented, such as:
    - A general evaluation framework, including both kinds of evaluation, such as the ISLE Framework for Evaluation in Machine Translation (FEMTI) approach
    - An integrated evaluation platform
    - In the same framework, remote evaluation distributed over the Internet, which permits to interchange components, allowing comparing various approaches, while also examining the influence of the component on the whole system, and which could be organized as Web services.

- Evaluation of the results of automatic techniques must also foresee complex scenarios where the quality of the final results depends on the quality of the partial results.

3.1.5 Directions for research and development

- Invest in the development of resources and technologies for processing non-verbal, and more generally contextual information encompassed in speech-based interaction
- As many of the automatic evaluation measures in the style of BLUE and its descendant are still highly controversial, active research into other types of metrics and other ways of evaluating is desirable
- More efforts are needed to solve the problem of how to automate the production of the large quantity of resources required, and at the same time how to ensure the necessary quality to get acceptable results in industrial environments
- Standards need to *co-evolve* at high speed together with rapid change in science, technology, commerce

3.2 Funding Agencies and Policy Makers

3.2.1 Resource production and use

- Use of best practices or standards in new projects must be enforced, to facilitate data re-use. Projects developing LRs should be requested to adhere to standards for encoding and representation of data and associated annotations
- Enforce documentation of resources, including annotation formats
- Priorities in the development of core LRs for languages should be driven by BLARK-like initiatives: support them and encourage countries to develop their own BLARK matrices
- The creation of LRs must be tied to the development of technologies. It is mandatory to produce the basic tools to process the 'raw' data
- Support the development of LRs for less-resourced languages
- Invest in the production of parallel corpora in multiple languages
- Support the development of resources and technologies for processing non-verbal, and more generally contextual information encompassed in speech-based interaction
- Actual industrial needs have to be to addressed: information about whether the resources acquired are actually used or, the other way around, of what are the particular characteristics of the actually used resources, needs to be made public. The involvement of industries in the research on automatic methods must be supported
- Public procurement, especially at the EU level, should be used as one of the instruments to boost production and adoption of language technologies.

3.2.2 Interoperability issues

- Try to solve the "standard divide" by which a few languages are very well equipped with language resources and consequently with LR standards needed
- Have an *integrative view* on LR standards: an European Interoperability Framework (EIF) for LR has to be developed (cross-domain, cross-purpose, cross-cultural, etc.)
- Contribute to expand the EIF, e.g. in the context of eGovernment, eHealth, eLearning, etc. where many of the existing LR standards can already contribute effectively to enhance data interoperability
- Bring together research communities and industrial application communities for developing a joint vision on LR standards in general
- Foster cooperation between MT industry and CAT-oriented translation and localization industry, for well-balanced and more integrative LR standards industrially usable yet based on pre-normative research
- Develop a broader vision of LR standards with the inherent inclusion of *multimedia*, *multimodal* and *speech* processing applications
- Create an operational ecology of language resource standards that are easily accessible, re-usable, effective, and that contribute to semantic interoperability
- Aim to a global standardization effort on the well-known line of EAGLES-LIRICS-ISO, a long-term strategy which brings together US-experts with their standards and best practices with the European traditions of EAGLES etc. and with East Asian best practices in the field.

3.2.3 Availability, licensing, maintenance and preservation

- It is important to ensure that publicly funded resources are made publicly available at very fair conditions. Public agencies should impose that resources produced with their financial support are made available free of charge for academic R&D activities. It is also important to encourage language resource owners to donate them to data centres to be distributed free of charge
- Enforce/sustain initiatives for data archiving and preservation: it should be ensured that the data produced by a certain project/initiative/organisation will survive any possible change of media for distribution
- When funding new LRs, request a plan for their maintenance
- Ensure sustainability of funded resources, e.g. by requesting accessibility and usability of resources for a given time frame
- Sustain initiatives offering legal recommendations/guidelines for the reuse of Language Resources, and investigating appropriate licensing models allowing for re-use and re-distribution.

3.2.4 Evaluation and validation

- The definition of appropriate evaluation frameworks for automatic acquisition methods is needed. The development of evaluation methods that cover the different automatic techniques is fundamental, in order to allow for a better testing of existing and newly discovered methods. Beyond the evaluation on scientific grounds, it is also recommended that techniques are measured by their impact in real scenarios of NLP applications

– Promote a permanent effort framework to take care of language technology evaluation in Europe.

3.2.5 Directions for industrial research and development

– Support the involvement of industries in the research on automatic methods, so as to allow a more precise assessment and evaluation of automatic methods for the development of LRs for real-scale applications
– Support transfer of Human Language Technology to SMEs: instruments should be established to transfer language technologies from projects to the SME language technology community in order to stimulate the availability of new technologies and increase the language coverage.

3.2.6 Policy issues

– New languages that joined recently the Union should be considered as a higher priority in coming EU programs
– Human language resources need to be "de-globalized" and focus on local languages and cultures despite today's "global" village
– Copyright law should be harmonised at the European and national level in such a way to permit the free use of copyrighted works for academic purposes[1]
– Favour multidisciplinary integration of different communities.

4 Recommendations for Infrastructures in the HLT field

4.1 The main message: an Open Resource Infrastructure to boost the Language Technology field

Infrastructural issues – such as interoperability, resource sharing, easy access to LR&T – were recurring messages in all the sessions at the FLaReNet Vienna Forum and in other FLaReNet or FLaReNet-related events. Many LR groups, initiatives and individuals have been advocating since some time the need of a language resource and technology infrastructure. This is now increasingly recognized as a necessary step for building on each other achievements, integrating resources and technologies and avoid dispersed or conflicting efforts. Infrastructure building is indicated as the most urgent issue: there is consensus on the fact that time is ripe for establishing an Open Resource Infrastructure (ORI), which allows easy sharing of data, corpora, language resources and tools that are made interoperable and work seamlessly together, as well as networking of language technology reserarchers, professionals, users. As a response to the community needs, FP7 Call 4 foresaw a Network of Excellence addressing such an infrastructure: the ORI will be implemented by the T4ME Net (Technologies for the Multilingual European Information Society) consortium.

1 By this recommendation FLaReNet endorses the document presented at the NEERI Conference in Helsinki (1-2 October 2009) "Freedom of use of copyrighted works for academic purposes" (by Ville Oksanen), from which the text above is slightly adapted.

An important factor for the success of an ORI is the acceptance and the active involvement of the community: this has to be carefully prepared. FLaReNet is paving the way to the set up and functioning of such an infrastructure through a number of preparatory initiatives (started or planned within FLaReNet) that must be continued and strengthened during the establishment and the life of the ORI. As such, these constitute a set of recommended sub-goals that must be pursued by FLaReNet now and together with T4ME in the near future to lay down the basis for an ORI:

- **Community building**: this requirement must not be underestimated
    - ***sensitising the community***, by spreading awareness of the need and motivation for such an infrastructure;
    - ***mobilising the community*** also through their active participation in a number of collaborative initiatives (started or planned), such as the LREC Map; the repository of guidelines, best practices, etc; the wikipedia of LRs;
- **Activities around and for the ORI**
    - ***catalysing*** a number of ***activities*** around the major theme of the ORI, to be seen as "the major" endeavour of the LR community in the next future;
    - creating ***consensus*** around ***interoperability*** issues and promoting standardisation activities (e.g. the planned "Repository of Guidelines");
    - providing easy and uniform ***access to*** the main available ***catalogues*** of LR&Ts;
    - fostering the knowledge dissemination and availability of language resources and tools that will eventually enter the ORI (e.g. through the "***LREC Map***");
    - promoting the creation of new language resources and the enhancement of existing ones;
- **Information gathering**
    - setting up ***expert groups*** that will work in synergy on a number of different, yet concurring issues crucial for the design and development of a language resource infrastructure, such as a preliminary assessment of the temporal, technical, organisational, and legal constraints involved in the implementation of an ORI;
    - ***eliciting user needs and requirements*** on a range of issues, from technical to organisational, related to the functioning of the ORI;
- **Cooperation**
    - ***boosting cooperation*** between existing infrastructures and initiatives;
- **Basic principles for the ORI**
    - pointing at a set of ***basic principles*** that will be at the basis of the design of an ORI and giving a first assessment of the various constraints involved in the implementation of an open resource infrastructure, such as, for instance:
        - **Temporal**: what is the time frame under which a first implementation can be reasonably expected (considering various steps and building blocks)?
        - **Technical**: which should be the main principles underlying its architecture, assuming some evolution over time to account for new developments (web2.0 …)?

- **Organisational**: Which are the best organisational and governance models that can ensure efficiency and sustainability? How the involvement of members should be regulated?
- **Legal**: what are the legal and IPR issues that must be addressed for the realization of an ORI?

A "first" set of basic principles and characteristics for the ORI – as emerging from a set of "concrete scenarios of use of the ORI" anticipated by the FLaReNet Steering Committee – is given in the following Table. These principles, still in a conception phase, must be exposed to all the community, so that many can participate in the discussion and a consensus is gradually formed around them. Some characteristics are accompanied by a first set of recommended actions.

Some **Basic Characteristics of the ORI**	**Related Actions**
Main purpose: Provide an ***environment for HLT Research and Development***	
Main feature: A ***community built and community maintained*** infrastructure, not imposed but emerging from and responding to its requirements	Launch initiatives with active involvement of many groups. Create the conditions for the community being actively engaged in its construction
Main feature: Must become ***"THE" place to go to get both information and data***	Promote awareness of recognition of the ORI as best point of entrance, but mostly achieve this through early evidence of its well-functioning
Openness: to be understood as an intrinsic multidimensional characteristic taking into account dimensions related to technical basis, legal aspects, licencing possibilities, sharing/distribution and pricing policies, etc.	
Wide awareness: everybody must know about the ORI	Wide information campaign is needed: prepare the field, identify the communities, gather people together
Acceptability: the ORI must not be felt as imposed from top but as a spontaneous outcome	Community mobilisation and involvement in discussion of basic principles
Broad coverage of LRs and LTs	Increase knowledge of LRs and LTs; LREC Map
What kind of ***resources in the ORI***: not only data and tools, but also guidelines, standards, evaluation packages, …	Start collecting information on all these, LREC Map
Common Platform for Distributed Resources: actual metadata and actual resources are distributed over the various different organisations/members of the ORI	Promote discussion on this issue, identify best possible appropriate architectures

Easy and Flexible access: Easy access to LR&T through central inventories and single sign-on. Access to actual data and tools can be either via download or via web services	
Service-based scenario: technologies and resources will be made dynamically available and accessible as a service through the net	Establish alliances of many resource providers
Simplicity: sale, licensing and clearance of IPR must be kept as simple as possible	Create a think-tank on these topics to start assessing requirements. Investigate and increase knowledge about existing legal schemes, their suitability for the LR&T domain
Profiling: every user needs to be registered to and authenticated/recognised by the ORI	See also CLARIN practices
Different profiles: providers vs. users, and different categories inside profiles (according, for instance, to the particular agreement providers propose for their sharing)	
Differentiation of access rights according to different profiles	
A unique point of access: the ORI is accessible through one common interface (residing on one or more servers) and aggregates (through harvesting) information and descriptions of resources from many different catalogues, in a transparent way for the user	Agreement is needed among major information and data providers, to create links among their information catalogues and data repositories. FLaReNet will facilitate interaction and cooperation among them
Persistent identifiers: links to resources must be maintained valid	See CLARIN solutions
Metadata must be harmonised and or mappings created across various catalogues. Metadata of different granularity may be needed	Initiate the process and bring it to a good level of development. Create a task force involving people from ISO, CLARIN, ELRA, LDC, OLAC, … Different granularity refers to e.g. browsing the catalogue(s) vs. analysing the details of individual resources/tools before using them
Tools to assist in metadata creation for new resources entering the ORI	
Provide ***recommendations on standards*** (and associated guidelines, best practices, schemata, etc.) and facilitate their use/adoption	Closely interact with SILT, CLARIN, ISO, W3C and other initiatives world-wide. Start collecting guidelines and schemata in an easily browsable repository, together with evaluation of standards and advantages in using them

Provide ***technologies for*** the creation, validation and conversion of new LRs according to the ORI ***standards***	Prepare through dissemination of best practices for major types of LR. Provide converters to/from main standards and best practice tagsets. Create tools to annotate in accordance with the major standardised formats
Auto-enrichment of the ORI: the results of using data or tools from the ORI – be they new or improved data or tools – must be made available afterwards on the ORI itself. This way the ORI auto-increases itself	This reinforces the notion of the ORI as a "community-built" infrastructure, in the way of a Wikipedia
Sustainability	Create a think-tank to assess requirements
Quality assessment	Create a think-tank to assess requirements. Give users the possibility to provide comments and votes, both for resources/tools and for standards.
Evaluation packages: the ORI should offer users of LT (and integrators) best practices and methodologies to evaluate a technology before acquisition	Analyse the evaluation packages that would be needed and those that are available
User-oriented interface and functionalities, such as: single sign-on login, resources/tools/services/ workflows overviews and (facetted) browsing via one entry point, predefined packages for frequently asked queries, licenses agreement, converters from the ORI standard formats to User's own formats, etc.	Design "web shop like" interfaces
Scalability of tools available on the ORI	
Legal aspects	See "Simplicity" above
Flexible availability wrt Fees: Data and tools can be accessed either freely or for a fee	
Business model: a service-based scenario, not based on licensing but on use	Stakeholders can and should participate in the definition of a business model
Trustability: trust is at the basis of the well functioning of the ORI; its success is determined by its use, and use depends on trustability	Liaise with major centres and data providers, etc. Provide a quality mark
Very broad platform of resource Providers and Users: Providers and users (institutional or individual) must "belong" to and must be recognised by the ORI	FLaReNet starts creating a large Network of providers and users (individual and insttitutional)

ORI should also act as an ***Observatory***, by offering services such as detection of gaps to funding agencies and programmes/project managers	
Governance of the ORI must be the outcome of collective and extensive involvement of the community	Involve the community in preparatory steps for the definition of governance model in all its dimensions. Ask for feedback on first draft of basic principles

4.2 HLT Stakeholders

We list in the following some general recommendations that can be used for an ORI, coming from the various FLaReNet events, presented in two sets targeted respectively to stakeholders and funding agencies.

4.2.1 Infrastructures for information and resource sharing

- Implement services and policies to enable sharing of resources
- Adopt a model for tool and resource development based on open advancement and collaborative development, where the community as a whole contributes components, modules, etc. to a common system or framework
- Another strategy proposed and partially implemented is to set up cooperative services by multiple local players to counter one big global player. Since each local player can often offer better quality than the global player for the particular local language or domain, the cooperating local players together can compete against the global player and offer customers the best available quality
- Ways of cooperation and linking among (partially similar) infrastructural initiatives both in Europe and in other continents must be devised and basic principles commonly agreed
- Models for giving credit for data sharing should be devised
- It is necessary to develop models and paths for turning research results into easily shareable results
- Initiatives such as the LREC Map of Language Resources and Technologies should be enforced and widely spread to the entire community, as they can turn into useful measuring tools for monitoring the evolution of LRs over time as well as to gradually lead to a situation where the notion of citation & publication of LRs becomes accepted and gives academic credit
- Since the lack of documentation and clear information about resources and related technologies is an important issue, harmonization of metadata and, at the same time, enforcing use of a common vocabulary of categories to describe and document resources are important steps towards facilitating surveying and retrieval activities.

4.3 Funding Agencies and Policy Makers

4.3.1 Infrastructures for resource sharing

- Improve access to digital research data for a better exploitation
- Data generated by public sector institutions should be increasingly made available for research and development, following principles similar to the Public Sector Information Directive
- Language resources built in the framework of EU or other funded projects should be made available at fair cost
- Infrastructure building is the most urgent issue. Infrastructures and repositories for tools and language data, but also for information on data (documentation, manuals, metadata, etc.) should be established that are universally and easily accessible by everyone
- In the long term, interoperability will be the cornerstone of a global network of language processing capabilities. The necessary framework and a corresponding infrastructure (i.e. standards and technologies) must be established and made operational. This can only be achieved through a coordinated, community-wide effort that will ensure both comprehensive coverage and widespread acceptance. Not only are data formats to be standardised, but also metadata
- An Open Resource Infrastructure should be established, which allows easy sharing of data, corpora, language resources and tools that are made interoperable and work seamlessly together
- An infrastructure for collecting data is needed. An appeal was made to the EC to support an infrastructure and tools to collect language data for a wide range of applications, as well as for the creation of data, in particular conversational speech data for speech-to-text technology for the whole range of European languages, and make these data available at affordable prices for research purposes and to SMEs
- Different funding agencies should jointly take care of infrastructural priorities
- Enhance current coordination of language resource collection between all involved agencies and ensure efficiency (e.g. through interoperability)
- Repositories of data formats, annotations, and guidelines should be supported as a major help to achieve and promote standardisation
- Since cooperation cannot be limited to a European landscape, alliances and cooperation among (partially similar) infrastructural initiatives both in Europe and in other continents must be organised and favoured.