www.meta-net.eu
office@meta-net.eu

META-SHARE Charter

Language Resources Sharing

Research and development in Language Technologies requires the availability of Language Resources (LRs) of good quality and quantity. Providing such LRs for all languages and for all technologies is very costly. LRs become valuable through sharing and cost-effective through re-using.

Clear, standardised and open terms of use for LRs reduce transaction costs and allow for their maximum utilisation. Less time has to be spent on legal clearance.

LRs should be open to use, reuse, sharing, improvement and deployment in order to advance research and foster development.

The META-SHARE language resource exchange facility is devoted to the sustainable sharing and dissemination of LRs between its members and the community at large, increasing access to LRs at a global scale, among others, by supporting the harmonisation of the laws governing copyright exceptions and limitations, fair use and fair dealing.

This document has been compiled within the Network of Excellence 'Multilingual Europe Technology Alliance (META-NET), co-funded by the 7th Framework Programme of the European Commission through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893), and META-NORD (grant agreement no.: 270899).

Version 1.03, May 2011

Background

Research and development in Language Technologies require (a) the availability of Language Resources, i.e. language data sets and basic language processing tools, of good quality and sufficient quantity, and (b) evaluation means allowing performance measurements. While essential, providing such resources for all languages, or for all language pairs in case of translation systems, and providing evaluation means for all technologies in all languages or language pairs, is very costly.

Language Resources constitute an essential infrastructure not merely for language related research but also for the development of language technology applications, products and services. It is, therefore, necessary to devise processes and rules that allow the widest possible dissemination, distribution and re-use of such resources. Re-use and re-purposing of Language Resources is an integral and essential part of the language technology development cycle and has to be taken into account in any related policy decision.

Language Resources may be of collective authorship. In addition, depending on the nationality of the author and the jurisdiction in which the LR is used, different laws may apply.

The existing legal system, particularly copyright law, does not follow a uniform approach for all types of language resources, containing different regulations regarding the treatment of text as compared to the treatment of audio, images or video and again different for software, technological applications and non-creative material in general. Furthermore, the rules regarding collective works, databases and works of shared authorship are not uniform across jurisdictions.

In addition, there is no international solution to the problem (a) of orphan works and (b) the use of Language Resources for the creation of (or contained within) language tools and technologies; it is not always clear whether specific licensing arrangements are required or whether the use of Language Resources is covered by copyright limitations and exceptions or fair dealing provisions. These are not harmonised either at the international or at the European level causing further uncertainty between language resources users.

Aim

The aim of this Charter is to give a clear motivation to Language Resource providers and users, the language technology community, market players, policy makers and the public so that in the digital world Language Resources be shared and further re-used with the minimum possible transaction costs and efforts and under clear and easy to understand rules. META-SHARE fully endorses and supports this vision.

General

The term Language Resources (LRs) in this Charter means all language-related digital assets including, without limitation, raw data, processed data, metadata and any other kind of data sets as well as language processing tools, technologies and language related services.

Open Source is used in this Charter in the sense of the Free Software Foundation (FSF) and Open Source Initiative (OSI) definitions; Open Content and Data are defined in accordance to the Open Knowledge Foundation (OKF) principles. Shared source, content or data are resources which follow the FSF, OSI and OKF definitions and are available to a group of specified or specifiable individuals or organisations.

Organisations or individuals working on LRs shall be encouraged to adhere to the guidelines described below. Policy makers should actively encourage them to follow these guidelines by providing funding, setting rules, giving instructions or by taking any other appropriate measure.

When managing LRs, all entities involved shall strive to promote and be in accordance to this Charter, the Panton Principles (http://pantonprinciples.org/), the Europeana Public Domain Charter (http://www.version1.europeana.eu/web/europeana-project/publications) and the COMMUNIA Public Domain manifesto (http://publicdomainmanifesto.org/).

Guidelines

1. Infrastructure and metadata

LRs have to be described with agreed metadata, which must be standardized and ideally open to fair harvesting. LRs should be persistently stored in an open and documented format.

Metadata of LRs should be visible, shared, open to re-use and indexed in such a way to enable LR effective search and discovery.

2. Standardisation and interoperability

Open Standards and best practices should be used for language data sets, processing tools and metadata, if they are available. Data formats have to be standardized and ideally open.

Software has to use open interfaces and be ideally Open Source.

3. Data sets and tools

Data sets and tools shall be ideally open or shared. If data-sets and tools are provided under commercial licences, these licences have to be standardized. The licences should ideally be readable and stipulate that the data are provided under non-discriminatory terms, through an on-line service or other low transaction cost mechanism.

LR creation should be given proper academic recognition, statistics should be collected with regards to the use of LRs and steps should be taken to introduce a “LR impact factor”.

4. Public Domain

Content and LRs that are part of the Public Domain should be clearly marked as such and must not be burdened with any additional restrictions.

5. Rights Clearance and Licensing

All necessary Intellectual Property Rights have to be cleared before any LR is disseminated or made available in any possible way or means to the public.

Any grant of access to LRs should include at least the right to both humans and machines to read the relevant content. Allowing transformative uses, dissemination and distribution of such resources is strongly encouraged.

To limit the complexity of licensing, standard, easy to use licences or a range of recommended licence templates or model licences have to be used.

If a LR is provided under copyleft (share alike, (SA)) conditions, i.e. if the LR's derivatives have to be shared under the same terms and conditions, such conditions must be part of a standard licence. A collection of LRs has to be licensed preferably under compatible SA licences.

6. Restrictions

The restrictions to use or re-use LRs should be the minimum possible so that (a) LRs are continuously enriched and (b) language technologies and related services are fully deployed and further developed.

If attribution is required, sufficient information to identify the attributed entity including names, nationalities and legal status shall accompany the relevant LR.

7. Data Protection and other Third Party Rights

As far as the material contains personal, sensitive or confidential data and as far as this cannot be countered by pre-processing the material, all reasonable efforts have to be made to obtain consent for the maximum possible use, re-use, dissemination and distribution of the material, at best at the point of collection, but in any case before sharing the LR or making it available to the public.

Effective privacy policies and tools facilitating anonymisation and consent management have to be put in place.

8. Open Market provisions

While it is encouraged that LRs be made available in an as open way as possible, any LR infrastructure should support all possible business models without imposing any restrictions or barriers to market entry.

If non-commercial obligations are attached to LRs, there has to be a precise definition of what constitutes commercial use as well as a clear, standardised and pre-defined way to obtain commercial rights.

9. LR Open Services

LR service providers should allow open transfer (portability), i.e. technically and legally unhindered collection and reproduction of the data upon which they are based and provide clear and open exit policies to their customers. This means that the users of a language service should be able to change service providers at zero or minimal cost and be informed about any conditions such change entails clearly and in advance.

10. Public Funding

If LRs are produced entirely with public funding, they have to be open or shared at least for research purposes. Such LRs have to contain clear attribution and rights-holders information, be properly documented and be made available with an appropriate licence, either as open or as shared with a fee that should not exceed the cost of their maintenance.