# Legal Issues for

# Sharing Language Resources:

# Constraints and Best Practices

Held in conjunction with the 7th International Language Resources and Evaluation Conference (LREC 2010)

17 May 2010

# PRESENTATION

## Organising Committee

- Khalid Choukri (Evaluations and Language resources Distribution Agency (ELDA), France)
- Denise DiPersio (Linguistic Data Consortium, USA)
- Marc Kupietz (Institut für Deutsche Sprache, Germany)
- Valérie Mapelli (Evaluations and Language resources Distribution Agency (ELDA), France)

## Description

Nowadays, the need for using and sharing Language Resources has become obvious to allow the development and improvement of Human Language Technologies. Upstream, and prior to any possibility of sharing, many legal issues have to be taken into consideration and solved. This goes from clearing IPR and data protection issues to defining licensing modes, etc. all along a Language Resource life cycle - collection, sharing with project partners, (re-)distribution -. In this regard, a series of obstacles need to be considered from a legal point of view. These obstacles range from the lack of legal concern shown by some organisations when it comes to sharing Language Resources, to the multiplicity of IPR, data protection and liability issues to be taken into consideration, as well as to the considerable number of home-made consent and license models that are currently out there for the users, and where the latter are often inspired by software license models (e.g. GNU GPL, Creative Commons), up to the diversity of legal protection modes all over the world.

There exist several elements of comparison between already existing license models, among which the following standout:

- **Domains of application:**

One can observe that some licenses can apply directly to Language Resources, whereas others are inspired from library products, software distribution industry or, more fuzzily, from any type of works of the mind (IP).

- **Types of use:**

Limitations in the use of LRs are levelled within the different licenses depending on the groups of users that are targeted. Some licenses focus only on educational purposes, whereas others may allow research use only or may cover the possibility of integration in distributable/commercialized products. In that area, questions about institutional versus individual licenses can also be raised.

- **Modification issues:**

Results obtained from the LRs can be modified following constraints laid down in the license. Some licenses may allow the full modification and reuse of the LRs whereas others may not allow any modification at all or may only allow them under certain circumstances. For example, one LR may be used only within a research team for its specific research or even a specific project/activity (e.g. evaluation). Types of use and modification issues are usually very cross-linked.

- **Re-distribution:**

"Re-distribution right" is one of the core concepts in free licenses. For example, the Creative Commons model includes a "sharing" mention, which implies a free re-distribution of the source resources. Other licenses may not allow so much flexibility.

These legal issues do not only have an impact on the sharing and/or distribution of Language Resources but also on the many processes all around. Legal aspects are important, for

instance, to understand the ownership of the Language Resource specifications, in particular when these are derived from the customization of an existing product. Furthermore, the maintenance (corrections of bugs, improvements, updates, etc.), merging/fusion and even integration into new resources may also need to be treated with the appropriate/specific license or may be just left open to never-ending discussions.

Similar problems concern the Evaluation outcomes. The organisation of evaluation campaigns showed up the genesis of new types of rights to be considered. Not only the exchange of Language Resources is important here, but also the sharing of metrics and other tools, within so-called "evaluation packages". Moreover, complementary issues need to be looked at carefully, such as the rights for the dissemination of comparable results (cross-system comparisons), the anonymization/protection of participating systems, etc.

## Main issues

The aim of this workshop is to attract participants interested in this topic and to address questions on all related topics, including, but not limited to:

- Expectations from the HLT Community
- Demands, interests and desiderata from the linguistic point of view
- IPR, data protection and ethical issues around Language Resources
- Technical measures for dealing with legal restrictions
- Best practices
- International diversity
- Towards a uniform license model: existing differences and merging possibilities
- New ideas and cooperation orientations

## Why this workshop is of interest?

The importance of the existence of mediating data centers (middle man) in this area has been well proved for about 15-20 years. These centers have worked on the improvement of the licensing models, as well as on the overall dissemination process. The increasing availability of resources directly from the Web (copyrighted and not) and the idea of using the Web as a shared community data repository/work space are examples of how technology is creating a wind of change. In the case of the latter, data centers have been asked about sharing language resources in computing "clouds" and through virtual sites (like Amazon’s Mechanical Turk), uses that are not necessarily compatible with data provider agreements negotiated a decade or more ago. Nowadays, we are reaching a point of reflection time. Time has come to assess the work carried out until now and to think about the next orientations to be undertaken.

The proposed workshop aims at enlightening the often fuzzy knowledge around legal issues that have to be dealt with at each step of the production and dissemination of a Language Resource. It also aims at showing new lines of work in that field, as well as new possible cooperation topics. The workshop will also be a good opportunity for all interested parties to air their views and have an open discussion.

**Proceedings will be published after the workshop.**