Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 321-322 recibido 29-04-10 revisado 18-05-10 aceptado 20-05-10

The Harvesting Day: una iniciativa para mejorar la visibilidad de los recursos lingüísticos

The Harvesting Day: an initiative to enhance the visibility of language resources

Carla Parra
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona, España
0034 935421193
carla.parra@upf.edu

Marta Villegas
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona, España
0034 935421193
marta.villegas@upf.edu

Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona, España
0034 935421193
nuria.bel@upf.edu

Resumen: *The Harvesting Day* es una iniciativa para garantizar la visibilidad, localización y descripción de los recursos lingüísticos mediante un conjunto básico de metadatos. Esta iniciativa aboga por un cambio de estrategia en el que los proveedores de recursos y tecnologías lingüísticos se convierten en responsables de la visibilidad de sus propios recursos así como de su documentación. Una vez creadas y almacenadas debidamente las descripciones de los diferentes recursos, los metadatos son recopilados de manera automática y periódica y se envían a los principales repositorios y catálogos virtuales garantizando así la visibilidad de los recursos así como la veracidad de sus datos, que de este modo se mantendrán actualizados.
Palabras clave: Recursos lingüísticos, tecnologías lingüísticas, descripciones de metadatos

Abstract: The Harvesting Day is an initiative to ensure the visibility, accessibility and description of language resources by means of a basic and metadata schema. This initiative believes in a change of strategy: resource and technology providers must be aware of the importance of ensuring the visibility of their resources, as well as the documentation thereof. Once language resources descriptions are appropriately created and saved, the corresponding metadata are automatically and periodically harvested and sent to the main virtual repositories and catalogues. This guarantees not only the visibility of language resources and technologies, but also the trustability of their data, which in turn is continuously updated.
Keywords: Language resources, language technologies, metadata descriptions

1 Introducción

La iniciativa *The Harvesting Day* tiene como objetivo permitir a los principales catálogos y observatorios de recursos y tecnologías lingüísticos recopilar automáticamente las características principales de los recursos y tecnologías lingüísticos garantizando al mismo tiempo que dichos datos siempre permanecen actualizados gracias a una rutina periódica de recopilación de metadatos. Para facilitar y homogeneizar esta tarea, los recursos y tecnologías lingüísticos se describen mediante una descripción mínima de metadatos (BAMDES: BAsic Metadata DEScription) que garantiza que todos los recursos se describen utilizando una serie de atributos comunes a todos ellos.

En el pasado, iniciativas como la del proyecto ENABLER ya abordaron la descripción de recursos lingüísticos con un esquema de metadatos común y de hecho nuestra propuesta de metadatos mínimos está basada en su estudio de todas las iniciativas de descripción de recursos y metadatos propuestas hasta el año 2003 así como en su propuesta de esquema de metadatos para la descripción de los distintos tipos de recursos lingüísticos. Cabe recordar su Declaración por el acceso libre a los recursos lingüísticos (*Declaration on Open Access to Language Resources*), en la que ponen de manifiesto que pese al esfuerzo y dedicación de iniciativas como ENABLER la

ISSN 1135-5948

información sobre la existencia y naturaleza de la mayoría de los recursos lingüísticos es muy pobre y tan solo una pequeña fracción de los mismos es visible para los usuarios interesados[1].

Siete años después, la situación ha mejorado y existen iniciativas dedicadas a la recopilación de información sobre recursos y tecnologías como el catálogo universal de ELRA, el *Natural Language Software Registry* del DFKI o el *Virtual Language World* de CLARIN. Sin embargo, y a pesar del gran esfuerzo realizado por los principales catálogos y observatorios, los costes de mantenimiento y gestión de los mismos para garantizar que están permanentemente actualizados son muy elevados ya que los datos que necesitan recopilar suelen ser particularmente difíciles de localizar. Este mismo hecho queda constatado en el Deliverable D6.1a del proyecto FLaReNet:

The compilation of information for this first survey was harder than expected because of the ***lack of documentation*** *for most of the resources surveyed. Besides,* ***the availability of the resource itself is problematic****: Sometimes a resource found in one of the catalogues/repositories* ***is no longer available*** *or simply* ***impossible to be found****; sometimes it is* ***only possible to find a paper reporting on some aspects of it****; and, finally, sometimes* ***the information is distributed among different websites, documents or papers at conferences****. This made it really difficult to carry out an efficient and consistent study, as the information found is not always coherent (e.g. not every corpus specifies the number of words it has) and sometimes it even differs from the one found in different catalogues/repositories.*

Así, podemos concluir que pese a la existencia de propuestas claras y detalladas como la de ENABLER para la descripción de recursos lingüísticos, los proveedores de recursos lingüísticos no las utilizan y por tanto no existe una descripción homogénea y consistente de los mismos.

[1] "The work of many initiatives and surveys such as the one of the ENABLER Project show very clearly that the general information about the existence and the nature of most language resources is very poor. Only a small fraction of them is visible for interested users". (ENABLER Declaration, 2003).

Desde nuestro punto de vista, el modo de solucionar este problema implica un cambio de estrategia: los proveedores de recursos deben ser los responsables de garantizar el acceso y la visibilidad de sus recursos, así como de mantener los datos acerca de los mismos actualizados y accesibles en todo momento.

2 *The Harvesting Day*

La iniciativa que proponemos consiste en el empleo de un esquema básico de metadatos común para cada tipo de recurso y su recopilación y distribución de manera automática a los principales catálogos y observatorios.

La rutina de recopilación automática de los datos está basada en el protocolo OAI-PMH para recogida automática de metadatos. A través de la web de la iniciativa (www.TheHarvestingDay.eu), los usuarios pueden rellenar unos sencillos formularios de descripción de sus recursos que generan de forma automática las descripciones en XML que posteriormente serán recopiladas. En la misma web está también disponible un paquete que los proveedores de recursos lingüísticos deben instalar en sus servidores para convertirse en proveedores de metadatos y permitir la recopilación automática de sus descripciones. De este modo, cada vez que se fije una fecha de recogida de datos, el robot de recogida visitará el servidor del proveedor, comprobará si existen actualizaciones o nuevos recursos y facilitará toda la información actualizada a los principales catálogos y observatorios de recursos lingüísticos.

3 *Conclusión*

The Harvesting Day no es una iniciativa aislada, sino un proceso descentralizado que se repetirá periódicamente para garantizar que todos los datos recopilados están siempre actualizados, logrando así que la validez y fiabilidad de los datos se incremente.

La iniciativa aboga por un cambio de estrategia en la que los proveedores de recursos y tecnologías lingüísticos asumen un rol protagonista. Ellos son los responsables de que los recursos y tecnologías del futuro estén mejor documentados y sean visibles y por tanto deben garantizar que la información sobre sus recursos está actualizada y presente en los principales catálogos y observatorios. *The Harvesting Day* es el modo de garantizarlo.