2 HIDRA's Motivation and Role in the CADIAL Project

Neda Erceg, Maja Cvitaš

This chapter introduces HIDRA, an agency of the Government of the Republic of Croatia and describes its mission. It also gives an overview of HIDRA's motivation to take part in the CADIAL projects and explains its role.

1 HIDRA's mission

The Croatian Information Documentation Referral Agency – HIDRA as an agency of the Government of the Republic of Croatia was established to perform information, documentation and referral activities that assure availability of the official documentation of the Republic of Croatia and its promotion in homeland and abroad.

Also, HIDRA's task is to obtain information from official sources, both Croatian and foreign, assure its prompt processing according to international standards and provide access to them for all the users.

HIDRA also takes part in developing the information infrastructure of the Government and bodies of authority of the Republic of Croatia based on new technologies thus enabling the Government to come closer to the e-government paradigm.

The role of HIDRA in the CADIAL project covers both of these lines of its activities: by compiling a large document collection in digital form, making it available to interested parties and provide its intelligent content indexing and search engine.

2 The structure of the official documentation

The Croatian official documentation and information published and/or authored by state or local authority bodies or bodies with public authority that HIDRA provides by collecting, processing and disseminating information on different media is composed of:

- legal regulations (laws, international agreements, decrees, statutes, regulations, decisions etc.);
- documents that the bodies are obliged to publish in accordance with their establishment acts (plans and activity reports, statistics, maps, strategic plans, drafts and bill proposals, standards etc.);
- expert studies, bulletins, journals, translations of the international official documents and other editions;
- forms, registries, databases published on the official web sites (as virtual collections).

This collection of documents is administrated through a bibliographical database, including the document content indexing performed by a unique thesaurus. The content indexing enables easier and faster retrieval within the overwhelming abundance of textual information that has been produced by the bodies of public authorities on the daily basis. So as official sources of information are concerned, a uniform terminology contributes to better communication between official information creators and their consumers, i.e., citizens, to whom the information is intended for.

3 Usage of Eurovoc

For the purpose of efficient content indexing, HIDRA has coordinated the efforts of experts in various fields with lexicographers and provided the Croatian translation of Eurovoc thesaurus – a general multilingual thesaurus developed for content indexing of official documents of the European Communities (http://europa.eu/eurovoc). HIDRA is responsible for monitoring all the changes and amendments in the official version of Eurovoc, for its adaptation and publishing the official Croatian translation in paper and digital form, particularly on Internet through its site http://www.hidra.hr/eurovoc.

As a multidisciplinary thesaurus, Eurovoc acts as a common denominator among all the documentation systems of the European Union and

its member and accessing countries providing a structured and controlled multilingual vocabulary (encompassing 21 fields and 127 micro thesauri with around 6,000 terms) covering all areas of EU activity.

Eurovoc was first published in the mid eighties as a result of a joint project between the European Parliament and the Office for Official Publications of the European Communities resulting in a tool for contents processing and indexing their respective documents, texts published in the Official Journal, as well as those of the national parliaments of member states.

Today Eurovoc covers the official languages of the European Union, Croatian as EU accessing country and some other languages such as Albanian and Russian.

In order to assure the content indexing of official documentation of the Republic of Croatia at the national level, HIDRA is constantly adapting the Eurovoc with specific descriptors which encompass:

- standardised names of bodies of public authority;
- geographical names;
- names of political parties;
- general and thematic descriptors that are specific for content indexing of the Croatian official documentation.

Eurovoc together with the Croatian appendix functions as a very good basis for both, overall international and Croatian national, level of communication and cooperation.

The published terminology manuals also provide an example of the possible direction of the further development of the official terminology of the Republic of Croatia.

4 Beyond the manual content indexing

HIDRA has gained valuable experience in content processing of official documents and publications of the Republic of Croatia by using Eurovoc as primary documentalistic indexing tool on its own document collections. Having in mind the broadness and quantity of indexed official documentation from all areas, particularly legal documents that are produced in the process of harmonisation of the Croatian legal system to the legal system of European Union, several relevant conclusions could be met:

- the Eurovoc thesaurus provides a useful semantic framework for development of a general purpose thesaurus that is used for content indexing of the official documents and publications of the Republic of Croatia for the purpose of their storing and accessing;
- Content indexing of the official information sources demands an excellent knowledge of a wide range of different professions and areas;
- Content indexing also demands an excellent knowledge of the selected thesaurus that is used for the same areas;
- Beside all the knowledge that is needed, content indexing varies according to the person that is producing the indexing. Even the same person in different times could perform the indexing differently;
- In the time when a number of official documents published on Internet is growing rapidly, the need for their quick and qualitative content processing also grows, thus making the demands on documentalistic services higher;
- In order to spare time and additional efforts, it is would be indispensable to process (i.e., index) the document in the moment of its origin and to publish the descriptors together with the document.

Following this conclusions, instead of additional education in a wide circle of authors of the official documents – what would include their education about the Eurovoc, the national appendix and techniques of content indexing – the solution was found in building the *automatic indexing system* in the form of an intelligent computational system for indexing of publicly available official documentation in Croatian language with descriptors from the Eurovoc thesaurus.

This decision was also the Croatian answer to the invitation of Joint Research Centre (JRC) of European Commission which encouraged the Republic of Croatia in 2004, as a candidate country for EU, to join the efforts of overcoming language barriers within the process of EU enlargement.

5 Automatic indexing system

Having in mind the role of HIDRA as governmental agency that deals with the official documentation of the Republic of Croatia, the production of such a system there was a natural lieu for this task. HIDRA joined

forces with two faculties from the University of Zagreb: Faculty of Humanities and Social Sciences (Department of Linguistics) and Faculty of Electrical Engineering and Computing (Department of Electronics, Microelectronics, Computer and Intelligent Systems) providing the necessary expertise in the fields of language technologies, computing and machine learning. These institutions started the project AIDE which was later continued within the joint Flemish-Croatian project CADIAL.

The Eurovoc based automatic indexing system used on the place of documents and/or information origin means:

- Simplified usage of Eurovoc thesaurus without the need of previous expert knowledge;
- Sparing of human time and effort;
- Higher quality of content indexing since it should be done by the author of the text;
- Sparing of time in services for cataloguing and content indexing since the documents are already in digital form and indexed in the time and place of their origin;
- Enables the retrieval of documents in any phase of its life-cycle and by all bodies of public authorities since documents are already indexed from their origin.

In the process of building the automatic indexing system, besides enriching the Eurovoc thesaurus, HIDRA has also taken over the task of preparing and permanently maintaining the collection of indexed official documents that is used in the machine learning techniques, i.e., training of the automatic indexing system in order to improve the results.

6 Role in the CADIAL project

The final goal is to produce the publicly accessible collection of the official documents, primarily legal documents, of the Republic of Croatia that are formally and contentually processed and indexed. Using this document collection every citizen, without any expert knowledge, could issue a plain text query in Croatian and should expect the insight into content of any relevant document. Having in mind the multilinguality of Eurovoc thesaurus, it also guarantees transparency of the official documentation of the Republic of Croatia in an international context since it gives the opportunity to retrieve documents using other Eurovoc lan-

guages as query languages. In this respect this system could provide also cross-lingual document retrieval.

With this publicly and multilingually accessible document collection HIDRA is realising its primary mission and that is enabling the permanent access to the official documentation of the Republic of Croatia to interested users.

Also, the publicly available tool for content indexing of the official documents contributes to the information infrastructure of public authority bodies of the Republic of Croatia. It can be useful for faster and high-quality indexing of individual texts at the time and place of their origin, but at the same time, also for indexing of already existing document collections.