

February 2009

Relevance

CLARIN community

for all communities

# Concept Registry Service

## What is it?

Defining domain concepts together with the terms denoting them in the various languages in openly accessible and widely accepted registries will become crucial for improving semantic interoperability also across languages. Users that include references to registered concepts in the schemas defining classes of documents allow others to interpret them. Search engines for example can exploit the fact that two documents from different authors refer to the same registered concept even if the labels used may differ. Such mechanisms only work when the organization behind it can guarantee a long lifetime of the registry and a stability of the concept definitions. Therefore, all the work described in this document is being done in close collaboration with ISO - the International Standardization Organization.

It should be noted that on purpose ISO TC37 did not include relations between the registered concepts. Experience has shown that it is almost impossible to agree on relations, since they are to a wide extent dependent on theories, languages and usage scenarios. Therefore, relations need to be maintained outside of the concept registry.

## What is it for?

Users increasingly often want to

- carry out operations across resources from different researchers and projects, however, they are faced with huge hurdles to recognize and overcome the differences in concepts and terminology;

- re-use predefined concepts to simplify their task in resource creation by pointing to a registered concept or at least define a formal relation to a registered concept to indicate some sort of semantic overlap;
- allow search engines and other operations to exploit the references pointing to widely agreed concept and terminology registries.

In a world where cross-project and even cross-disciplinary work will become an increasing requirement researchers will be forced to make the semantic of the tags they are using more explicit. This can be done by referring to concepts registered in open and persistent domain registries.

## Who can use it?

- The ISO 12620 data model is meant to be generic, i.e. it is not limited to the linguistic community and should be useful for other disciplines as well. In the same way the ISOcat database software implementing the model and the decision processes can be used by other disciplines. The software will be freely available allowing anyone to create his own registration instance.
- The ISOcat service contains domain specific concepts (data categories) that may be of use for other disciplines as well. Everyone can refer to any registered concept in ISOcat and anyone can propose the registration of a new concept. A formal decision process according to the ISO rules has been defined.



## When can it be used?

In collaboration with ISO TC37 a data model has been worked out and it has been adopted as ISO 12620 standard. The software is being implemented as ISOcat registry service and will be launched in March 2009 for expert usage, and in summer 2009 for open usage, a wide variety of concepts emerging from earlier harmonization initiatives have already been added to be subject of the acceptance process. Contracts have been signed between ISO and the Max Planck Institute as official registration authority and the decision boards have been formed.

Since the ISO 12620 data model only specifies a flat list of concepts, a simple to use framework needs to be established to allow users to specify and manage relations amongst the various concepts.

## How does it work?

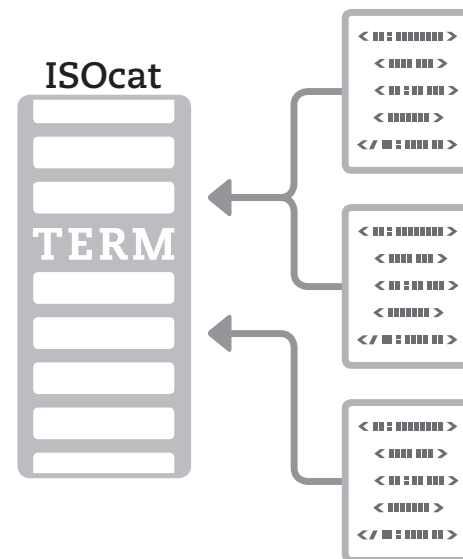
There are different strands that need to be considered: (1) Referring to a data category in ISOcat; (2) Suggesting a category for acceptance; (3) Establishing and using a relation; (4) Creating your own registry instance in case of a specific sub-discipline interest or to set up a concept registry for another discipline, for example.

(1) When users want to refer to a concept they simply look it up and use the identifier in conjunction with the ISOcat URI which will be maintained by ISO: <http://www.isocat.org/datcat/<dc>>. It is intended to have two mirror sites in the US and Korea for the ISOcat service to achieve a high availability.

(2) When someone wants to suggest a new concept to be registered or an existing one to be changed, he/she needs to register, create a new data category and then submit it for standardization

(3) Yet CLARIN cannot offer a handy mechanism for creating, manipulating, sharing and storing relations. This will be worked out in 2009. Currently, private mechanisms need to be supported.

(4) Yet the software is still in development with respect to some functions. Therefore we will not yet have a download option. We suggest getting in contact with the Registration Authority or the ISOcat service expert and asking for the exact state of the software and the conditions. The software will be offered open source and free of charge for academic users. Every concept registry service should be aware that a high level of availability and persistence will absolutely be required.



## Who is responsible?

CLARIN fully supports the ISOcat developments and will suggest concepts for the standardization process; however, it does not have an official role. Different groups are taking responsibilities:

- there are boards governing the concept registration process which take care of the ISOcat registry as a whole and for the different thematic profiles;
- there is the registration authority which is responsible for running the service, further developing the software and as a link to ISO;
- all is governed by the official ISO committees, in particular TC37/SC4

## Whom to contact?

The most important contact points are:

Decision Boards: Sue Ellen Wright, [sellenwright@gmail.com](mailto:sellenwright@gmail.com)

Registration Authority: Marc Kemps-Snijders  
[marc.kemps-snijders@mpi.nl](mailto:marc.kemps-snijders@mpi.nl)

ISOcat Service: Menzo Windhouwer  
[menzo.windhouwer@mpi.nl](mailto:menzo.windhouwer@mpi.nl)

ISO TC37/SC4 chair: Laurent Romary  
[Laurent.Romary@inria.fr](mailto:Laurent.Romary@inria.fr)

## Where to find more information?

The official ISOcat web-site is the source of all information:  
ISOcat: <http://www.isocat.org>

Responsible for the content:

**Marc Kemps-Snijders, Menzo Windhouwer**  
MPI for Psycholinguistics  
Wundtlaan 2, 6525 XD Nijmegen, NL  
Website: [www.isocat.org](http://www.isocat.org)  
Email: [marc.kemps-snijders@mpi.nl](mailto:marc.kemps-snijders@mpi.nl)