



ISO TC 37/SC4 N028
ISO TC 37/SC4/WG1 N10
2002-08-23

Language Resource Management
Descriptors and Mechanisms for Language Resources

File ID wg1N10.doc (1 757 ko)
wg1N10.pdf (355 ko) wg1N10.ppt (365 ko) wg1N10.zip (1 590 ko)

Title : A data category registry for language resources

Editor(s) : Sue Ellen Wright

Source : WG 1

Project number : N/A

Status : Contribution – Slide show presented during the Vienna meeting Austria 2002-08-19/23

Date : 2002-08-23

Agenda / Action : For consideration, discussion and comment

References : ISO 12620 (version under revision); ISO 16642

Mr. Key-Sun Choi - SC4 Secretary – KORTERM - 373-1 Kusong-Dong Yusong-gu - Taejon 305- 701 - Korea
82 42 869 35 25 – fax: 82 42 867 35 65 - kschoi@cs.kaist.ac.kr – <http://korterm.kaist.ac.kr>

Any question about SC4/WG1 documents registration, maintenance and distribution must be forwarded to Afnor
France preferably to: sylvie.arbouy@afnor.fr

Slide1

A Data Category Registry for Language Resources

Sue Ellen Wright
Kent State University
Institute of Applied Linguistics
© Sue Ellen Wright 2002

KENT STATE UNIVERSITY

1 of 29

Slide2

Data Category

data category
result of the specification of a given data field [ISO 1087-2:2000], i.e., a type of data field, such as *definition*

- Note: ISO 12620-2 is an inventory of data categories for use in terminology resources.
- Note: The term *data category* used in ISO TC 37 focuses on the type aspects of a group of defined data units of which the individual instantiations serve as tokens. *Data categories* in TC 37 usage correspond directly to *data elements* in the environment of ISO/IEC 11179.

KENT STATE UNIVERSITY

2 of 29

Slide3

Data Modeling Variance

- `<synonym>...content is a term that is a synonym of the preferred term or main entry term</synonym>`
- `<term>the preferred term goes here</term>`
- ...
- `<term>the synonym goes here</term>`
- `<termNote type="termType">synonym</term>`


KENT STATE UNIVERSITY

3 of 29

Slide4

Data Category Types

- **Complex data categories: have content**
 - **Open data categories: completely open content; any string can be the value**
 - **Closed data categories: predefined domain values (a picklist)**
- **Simple data categories: are content; the values assigned to closed data categories**
- **All treated in ISO 12620**




4 of 29

Slide5

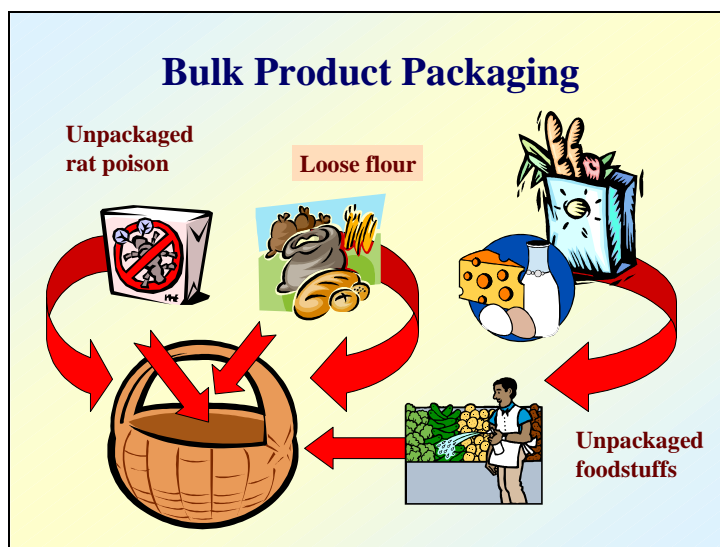
Data Categories: Packaging Data

- **Visualizing & managing information**
 - **An amorphous flow of undifferentiated stuff? vs.**
 - **An aggregate of individual elements (little packages, components) that can be identified, delimited, organized (modeled), stored, retrieved, manipulated, and reused**
 - **Stuff people can figure out if they think about it vs.**
 - **Stuff a computer program can automatically recognize and process**



5 of 29

Slide6



Slide7

03.02.1997 - 18:16:50 Walther 25.01.2000 - 18:56:54
 Wright 34 Network management alias 97/02/10 -
 10:29:28 Walther 97/02/10 - 10:29:28 Walther noun
 Main Entry Term shorter form of a long name, such as
 an email address, a directory, or a command Fahey:7
 Actually, all text-type addresses, long or short, are
 aliases of the IP address which is numerical only.
 Fahey:7 Aliases in real-time chat are usually
 referred to as nicknames and handles nickname
 noun Synonym When you start an IRC session, you
 specify a nickname, up to nine characters, which you
 can also change at any time. Osborne94: 413 handle
 noun The nickname you assign yourself when conversing
 in discussion groups. Falc? : 94 Use only in
 colloquial situations. alias noun m un nombre corto o
 apodo que se utiliza para enviar mensajes a aquellas
 direcciones de uso m? frecuente. **Unpackaged Data**
 alias representa a un grupo de per
 que se env? un correo a un alias
 todo el grupo de personas que lo forman. Carballar94:
 112

Slide8

The screenshot shows a software window titled 'Unpackaged Multi Term: EN/ES/ES/EN - IRC/MS - 1.1 MW - aliases'. It features a search bar with 'alias' entered and a 'Target' dropdown set to 'Español'. Below the search bar, there is a list of entries for the 'alias' category, including definitions in English and Spanish. A red box labeled 'Organized Data' highlights the search results area.

Slide9

data category specification


- listing of all the attributes associated with one given data category

The diagram consists of a vertical blue arrow pointing upwards and a horizontal blue arrow pointing to the right, both originating from a small box at the bottom left. This box contains the logo for 'KENT STATE UNIVERSITY'.

Slide10

Data Category Registry (DCR)

- all data category specifications used for language resource markup in all language resource domains (LRDs) of ISO TC 37: Terminology and other language resources


10 of 29

Slide11

DC Registry

Provides sample data category specifications from the DCR.


broader concept generic	
DC Name	broader concept generic
DC Definition	A concept (in a more level of abstraction higher than subject concept) in a generic hierarchical concept system.
DC Source	See ISO 10272000, 3.2.21 for a definition of "generic relation".
Comment	
Concept-related	Relations between concepts in generic systems are typically "is a" relations. This means that any given narrower concept "is an" instance of its superer broader concept (e.g. in a carter animal, and a carter animal is a mammal).
Example	In source C, figure C.5 and C.6, "mammal" and "lion" are broader concepts.
Data Type	plainText (open)
Target	ISO
Level(s)	Term Entry Language Section

term	
DC Name	term
DC Definition	A verbal designation of a general concept in a specific subject field.
DC Source	For definition of constraints, see ISO 1027-1, 1.4.1.
Comment	
Concept-related	Terms can consist of single words or be composed of combined string. The distinguishing characteristic of a term is that it is assigned to a single concept, as opposed to a pluriconceptual, which combines more than one concept in a hierarchical fashion to represent complex situations. Quality assurance systems is a term, whereas safety quality requirements in a pluriconceptual, specifically a collection.
Example	"truck" in source C, figure C.1.
Data Type	textText (open)
Level(s)	Term Entry Term Component Section

Slide12


Data Category Set (DCS)

- group of datcats selected for use in a specific LRD, e.g., ISO 12620-2
 - Note: Individual SCs and WGs within TC 37 can define their own Data Category Set (LRD-specific view of the DCR) by extracting datcats from the DCR (subsetting) and extending the DCR by adding specific datcats they need that are not already in the master DCR (supersetting).


12 of 29


Application Profile

- **subset of data categories selected from a DCS for use in a given language resource, e.g., for use in any given local application**
 - **Note: In cases where individual application profiles are designed for use across LRD boundaries, data categories can be selected from the DCR itself or from more than one DCS.**




13 of 29

Data Category Registry (DCR)




- **Core DatCat Repository for all TC 37 language resources**
 - **Conformant to ISO 12620-1**
 - **Containing all data categories used in TC 37 metadata registries**
 - **Available online in RDF format**
 - **Used for selecting LDR-specific Data Category Sets (DCS)**
 - **Used for configuring Application Profiles that go beyond individual DCS environments**



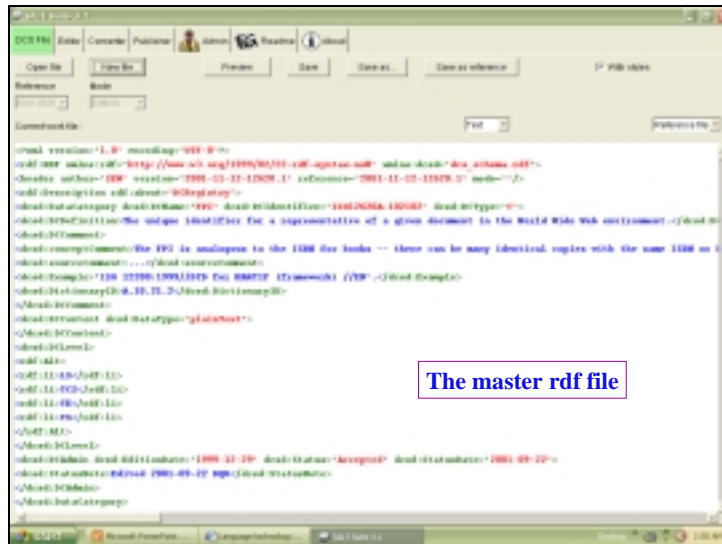
14 of 29

Implementation of the DCR

- **Based on the attributes defined in ISO 11179-3**
- **Augmented by additional attributes implementing data categories for language resources, e.g.:**
 - **Styles, vocabularies, annotation, refinement (in relation to ISO 16642)**
 - **Locale label**
- **Implemented and maintained as an RDF resource available online**



15 of 29



Rationale-1

- To provide implementers with means to design specific formats (in terminology, in other lexicography, in corpus analysis, etc.)
- To enable implementers in different domains to utilize common descriptors across LRD boundaries in order to ensure global interoperability
 - Example: terminological and lexicographical views of a single data set

17 of 29


Rationale-2

- To provide a global repository of all data categories used in TC 37
- To facilitate efficient delivery of a complete machine-readable (RDF) resource
- To enable the use of a single xml namespace for all TC 37 data categories

18 of 29

Namespaces


- **Note: bogus examples! Not real! These namespace files aren't really on the Web yet.**
- **ISO 12620**
 - **xmlns=http://www.iso.ch/iso/1999/ISO12620/iso12620**
- **OLIF**
 - **xmlns=http://www.bsi-global.com/2002/namespaces/OLIF**



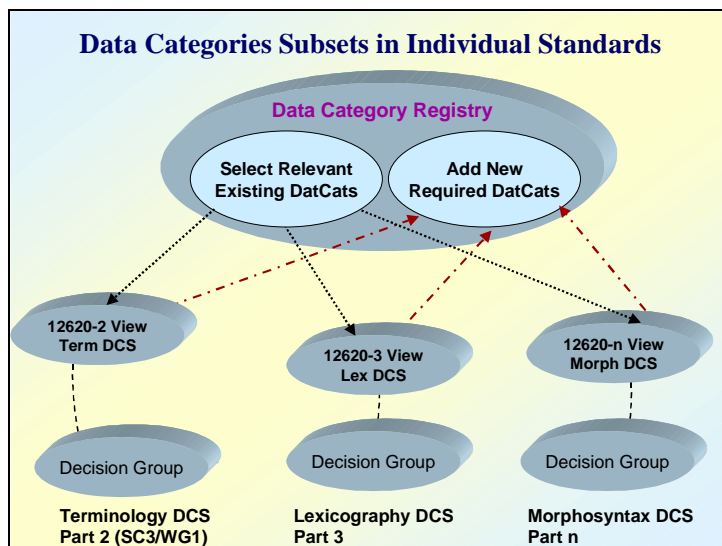
19 of 29

Implementing Namespace Capability

- **<context> in OLIF is not the same content from <descrip type="context"> in ISO 12620**
- **<OLIF-context> + <descrip type="ISO12620-context"> will preserve the integrity of the two datcats**




20 of 29



Slide22

Philosophy behind ISO 12620 Multiple Parts (Multiple DCSs)

- Thematic management of the data categories needed for individual language resource domains (LRD) resides in the Working Groups
 - A given WG selects the subset from the global set that is essential for that LRD
 - 12620-2 for terminology (SC3/WG1)
 - Potential 12620-3 for lexicography, etc.
 - The respective WG also specifies and datcats native to its domain




22 of 29

Slide23

Content of 12620-2, 3, 췌

- LRD-specific DCS comprises:
 - Selection of datcats from the global DCR
 - Additional constraints on content values
 - Constraints on the applicability of datcats with regard to the relevant metamodel
 - Constraints on refinements and annotations
 - Addition of required datcats not previously present in the DCR




23 of 29

Slide24

Content of Individual Standards

- LRD-specific view on the DCR
 - List of data category names required for a given LRD
 - Related data category definitions
- Other information associated with data categories (comments, examples, administrative information, ontological information, etc.) remains in the DCR



24 of 29

Slide25

DCRegistry
over time sample data category specifications published in a DCR standard (e.g., ISO 15926-2)

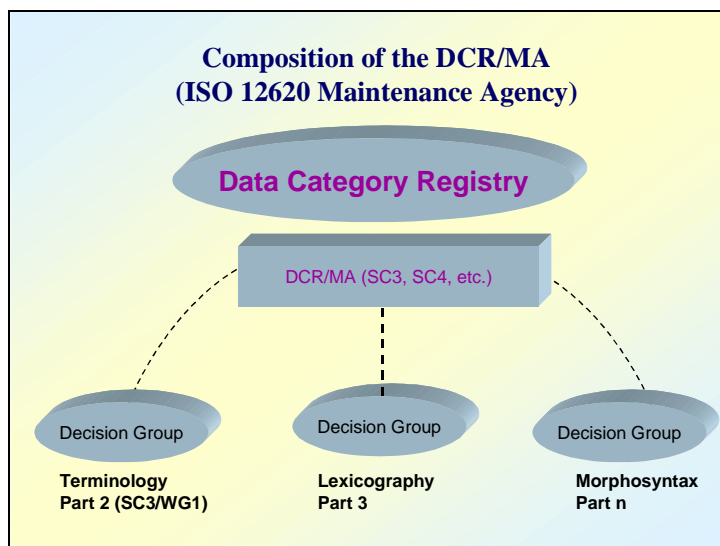
terminology concept generic	
DC Name	terminology concept generic
DC Definition	A concept from a more general level of abstraction higher than subject concept in a generic hierarchical concept system

term	
DC Name	term
DC Definition	A verbal designation of a general concept in a specific subject field

**Published standard lists DCName and DCDefinition.
 Electronic Web resource provides complete data set in RDF format.**

- Saves paper space
- Provides data in processable form.


Slide26

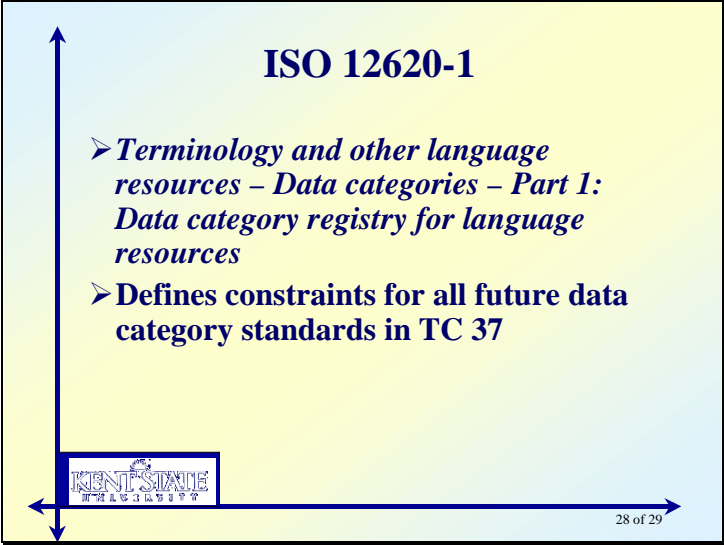


Slide27

Data Category Maintenance

- 1 maintenance agency for datcat registry for language resources
 - DCR/MA
 - Comprised of representatives from the various working groups
 - Plays a harmonization role
 - LRD-specific decision groups
 - Are comprised of representatives from the appropriate working groups
 - Select and approve data categories

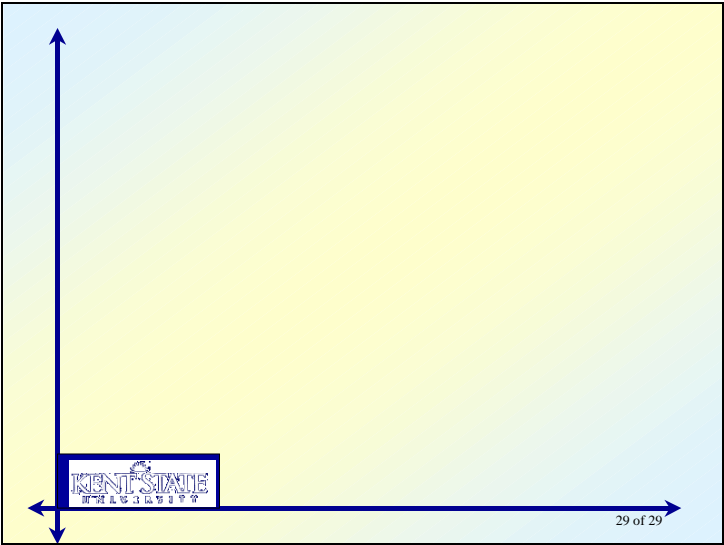
 27 of 29

A slide with a light blue to yellow gradient background. It features a vertical blue arrow on the left and a horizontal blue arrow at the bottom. In the bottom-left corner, there is a small logo for Kent State University. The text on the slide is as follows:

ISO 12620-1

- *Terminology and other language resources – Data categories – Part 1: Data category registry for language resources*
- **Defines constraints for all future data category standards in TC 37**

28 of 29

A slide with a light blue to yellow gradient background. It features a vertical blue arrow on the left and a horizontal blue arrow at the bottom. In the bottom-left corner, there is a small logo for Kent State University. The slide is otherwise empty of text.

29 of 29