The Well-Formed Data Category Specification

CRITERIA FOR STANDARDIZING DATA CATEGORIES

SUE ELLEN WRIGHT
METADATA TDG WEBINAR
2011-05-16

Data Category Names

- Lowercase (have been harmonized here)
- Status: *preferred name* (triggers the mode in the html “print” output)

capture method

PID: http://www.isocat.org/datcat/DC-2563

Identifier: captureMethod Type: complex/open Origin: CLARIN Profile: Metadata

Definition: Indication of the capturing/digitization method that was used when creating the digital version.
Source: CLARIN

Example: recorded digitally;digitized from VHS/Beta;digitized from tape/cassette;scanned/OCRed from paper
Source: CLARIN

Language sections: English, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, French, German, Greek, Hungaria

Data type: string

Preferred and Admitted Names

part of speech

Admitted DC Name

pos

PID: http://www.isocat.org/datcat/DC-396

Identifier: partOfSpeech Type: complex/closed Origin: ISO 12620 Profile: Terminology

Definition: A category assigned to a word based on its grammatical and semantic properties.
Source: ISO12620

Example: noun
Source: Mitre; TEI(green text); 1951

Language sections: English, German

Linguistic sections: German

Data type: string

- Use admitted term for alternate “synonym” DCs

Convert Value to Preferred Name

- TDG member notes need for change in the Submission Forum.
- Chair edits the DC Spec.
- In the Description Section, change the status of the English Name to *preferred name*.
- This definition is good.

Definition Form

- Superordinate concept + defining characteristics
  - **Indication** of the capturing/digitization method that was used when creating the digital version.
- Indication = superordinate element
- What kind of indication?
  - of the location where the tool/service is being executed.
- Capital letter & ends in a period
- Single sentence fragment
  - No extra sentences
  - No complete sentences
  - Avoid tautologies

Flawed Definitions

annotation workflow

PID: http://www.isocat.org/datcat/DC-2508

Identifier: annotationWorkflow Type: complex/open Origin: CLARIN Profile: Metadata

Definition: Indicates the workflow process in which the creation process was embedded.
Source: CLARIN

Definition: Indication of the workflow process in which the creation process was embedded.
Source: Based on CLARIN

Language sections: English, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, French, German, G

Data type: string

- Don't start a noun definition with a verb; use a comparable noun: Indicates → Indication of …

Flawed Definitions

character set

PID: http://www.isocat.org/datcat/DC-2565

Identifier: characterSet Type: complex/open Origin: CLARIN Profile: Metadata

Definition: The repertoire of characters used in the resource. A range of characters (non-coded character set) or a coded character set as defined in RFC 2050.
Source: CLARIN

Definition
The repertoire of characters used in the resource. A range of characters (non-coded character set) or a coded character set as defined in RFC 2050.
The repertoire of characters used in a resource, consisting of a range of characters (non-coded character set) or a coded character set as defined in IETF RFC 2050.

- Two sentence fragments; solution: combine to form a single definitive statement; identify RFC a little more clearly

Flawed Definitions

audio

PID: http://www.isocat.org/datcat/DC-2653

Identifier: audio Type: simple Origin: IMDI http://www.mpi.nl/IMDI/Schema/MediaFile-Type.xml Profile: Metadata

Definition: audio data. "Audio" requires an audio output device (such as a speaker or a telephone) to "display" the contents.
Source: http://www.ietf.org/rfc/rfc2046.txt

Definition: Any electronic media content that requires an audio output device (such as a speaker or a telephone) to "display" the contents.
Source: Based on http://www.ietf.org/rfc/rfc2046.txt

- 1) Fragmentary definition
- 2) Definition harmonized to reflect a set of closely related definitions for values of the data category *media type*.

Harmonizing Definitions for Picklist Values

- **media type**
  Definition: Specification of the media type of the resource or the media types for which the tool or service is suitable.
  - **audio**
    Definition: Any electronic media content that requires an audio output device (such as a speaker or a telephone) to "display" the contents.
  - **document**
    Definition: Any electronic media content (other than computer programs or system files) that is intended to be used in either electronic form or as printed output.
  - **drawing**
    Definition: Any electronic media content representing a picture, likeness, diagram or representation produced by marking lines as with a pencil, pen, crayon, or computerized drafting application.
  - **image**
    Definition: Any electronic media content that requires a display device (such as a graphical display, a graphics printer, or a FAX machine) to view the information.
  - **text**
    Definition: Any electronic media content consisting of a non-binary human-readable sequence of characters and the words they form that can be encoded into computer-readable formats such as ASCII or UNICODE.
  - **video**
    Definition: Any electronic media content that requires the capability to display moving images, typically involving specialized hardware and software.

Faulty Definitions

drawing

PID: http://www.isocat.org/datcat/DC-2657

Identifier: drawing Type: simple Origin: IMDI: http://www.mpi.nl/IMDI/Schema /MediaFile-Type.xml Profile: Metadata

Definition: A picture, likeness, diagram or representation
Source: http://en.wiktionary.org/wiki/drawing

Definition: Any electronic media content representing a picture, likeness, diagram or representation produced by marking lines as with a pencil, pen, crayon, or computerized drafting application.
Source: Based on Cambridge Dictionaries Online, "draw," "drawing"

- Misleading definition: some *pictures* are photographs and others are paintings; not all are drawings. Also: harmonization to match other values.

Near Tautology

text

Definition: any text-based information that is available in a digitally encoded human-readable format
Source: Wikipedia

Definition: Any electronic media content consisting of a non-binary human-readable sequence of characters and the words they form that can be encoded into computer-readable formats such as ASCII or UNICODE.
Source: Based on http://whatis.techtarget.com/definition/0,,sid9_gci213125,00.

- A text is text-based information – see revised definition; text is difficult to define.

Adjective Definitions

very good

PID: http://www.isocat.org/datcat/DC-2670

Identifier: veryGood Type: simple Origin: IMDI: session.resources.mediafile.quality Profile: Metadata

Definition: Pertaining to very good quality.
Source: CLARIN

Note: Corresponds to number 5 in IMDI format of Quality

- Change to:
- Definition: Pertaining to very good quality.
- General rule: Adjectives defined using *pertaining to, referring to, etc.*