Data Category specifications

What is a Data Category?

- The result of the specification of a given data field
  - *A data category is an elementary descriptor in a linguistic structure or an annotation scheme.*

- Specification consists of 3 main parts:
  - *Administrative part*
    - *Administration and identification*
  - *Descriptive part*
    - *Documentation in various working languages*
  - *Linguistic part*
    - *Conceptual domain(s for various object languages)*

Data Category example

- Data category: */Grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
  - Descriptive part:
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
    - French definition: Catégorie fondée (selon la langue) sur la distinction naturelle entre les sexes ou d'autres critères formels.
  - Linguistic part:
    - Morposyntax conceptual domain: */male/, /feminine/, /neuter/*
    - French conceptual domain: */male/, /feminine/*

Mandatory parts of the specification to be provided by the user

- For each data category:
  - – a mnemonic identifier
  - – an English definition
  - – an English name
- For complex data categories:
  - – a conceptual domain
- For standardization candidates:
  - – a profile (other then Private)
  - – a justification

Administrative Information Section

- The mandatory identifier
  - is a mnemonic string used to refer to the data category
  - should be based on a meaningful English word or series of words presented as an alphanumeric character string; for multiword strings, begin with lowercase and express the identifier as one continuous string in camel case with no white space (for instance, */term/*, */normativeAuthorization/*, */preferredTerm/*)
  - maybe used in XML vocabularies and thus must be a valid local part of a qualified name:
    - Cannot start with a number, shouldn't contain any whitespace, …

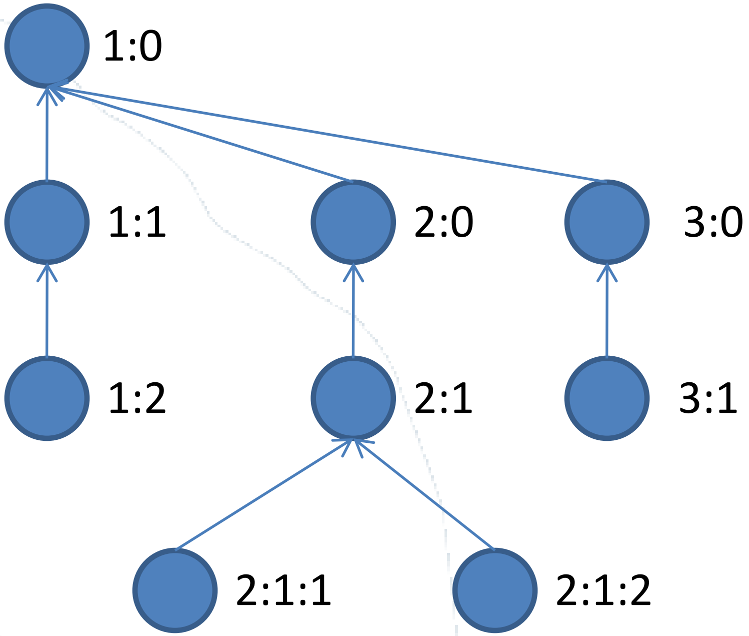
ISOcat warns you when the identifier is invalid and will refuse to save the data category

Administrative Information Section

- Each Data Category should be uniquely identifiable
  - Ambiguity: different domains use the same term but mean different ‘things’
  - Semantic rot: even in the same domain the meaning of a term changes over time
  - Persistence: for archived resources Data Category references should still be resolvable and point to the specification as it was at/close to time of creation
- Persistent IDentifier
  - ISO 24619:2011 Language resource management -- Persistent identification and access in language technology applications
  - ISOcat uses ‘cool URIs’
    - http://www.isocat.org/datcat/DC-1297 (/*grammaticalGender/*)
  - managed by the system

Administrative Information Section

- The mandatory version
  - used to refine identifier to indicate the version of the data category
  - managed by the system
  - based on branching not on major and minor revisions

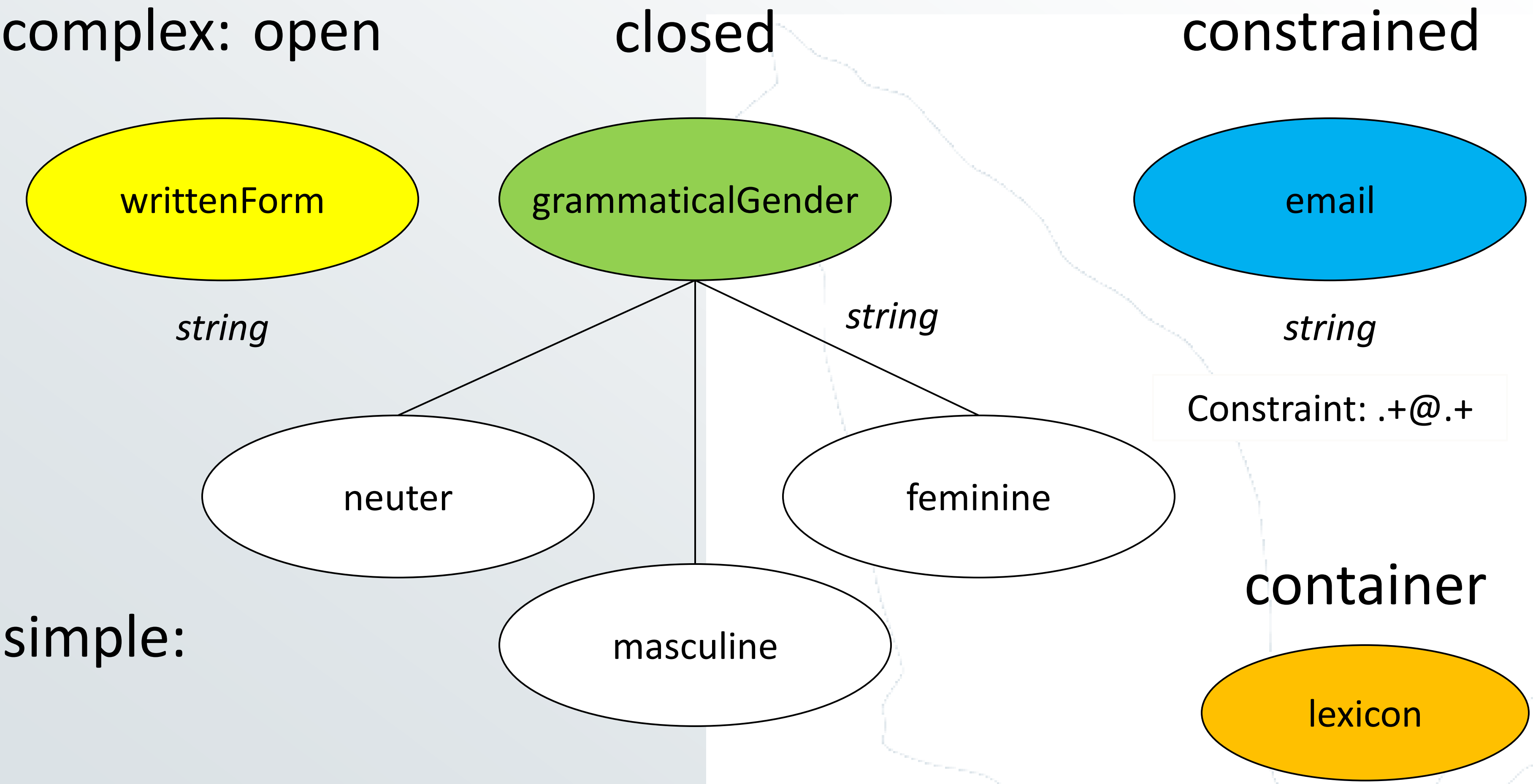


ISOcat version scheme will be revised

Administrative Information Section

- The mandatory data category type
  - container
  - complex/open
    - conceptual domain is not restricted to an enumerated set of values
  - complex/constrained
    - conceptual domain is non-enumerated, but is restricted to a constraint specified in a schema-specific language or languages
  - complex/closed
    - conceptual domain is restricted to a set of enumerated simple data categories making up its value domain
  - simple
    - describes a possible value in the value domain of a closed complex data category

Data Category types



complex: open

closed

constrained

writtenForm

string

grammaticalGender

string

neuter

masculine

feminine

email

string

Constraint: .+@.+

simple:

container

lexicon

Administrative Information Section

- The justification
  - a short description justifying why the data category should be included in the registry
  - mandatory for data categories to be standardized; desirable in general
  - even data categories that are common in a given thematic domain may be unfamiliar or ambiguous to users unfamiliar with that domain

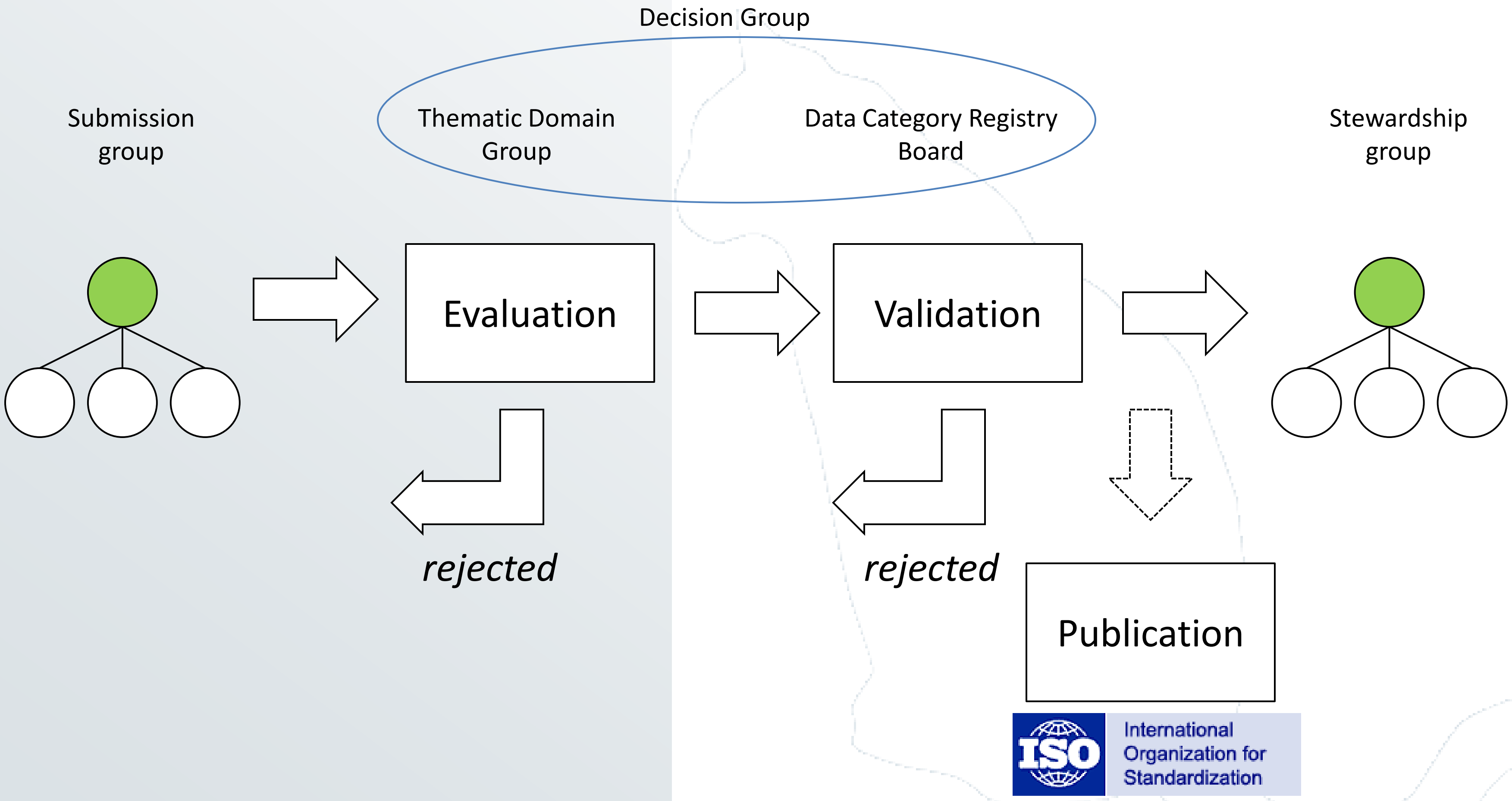
Administrative Information Section

- The origin
  - (document, project, discipline or model) for the data category specification

Administrative Information Section

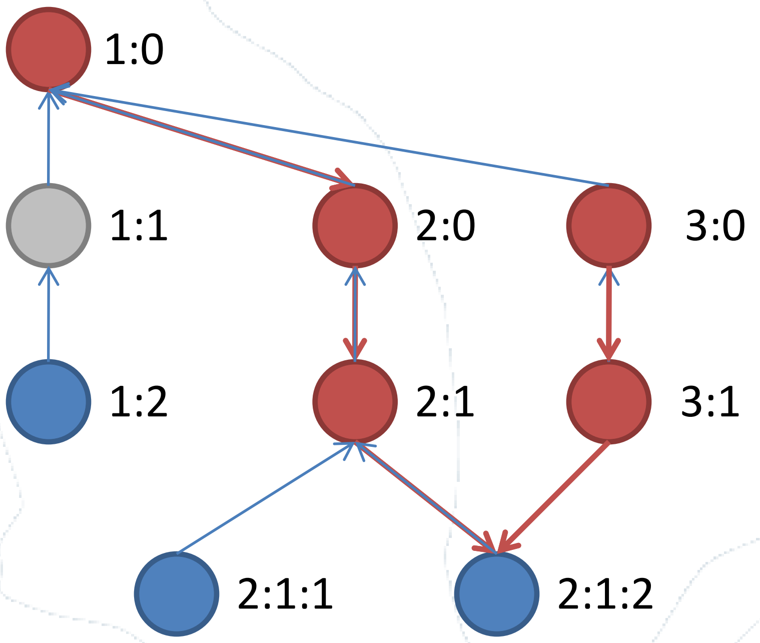
- The mandatory administration status
  - a designation of the status in the administration process for handling registration requests under the stewardship of the DCR Board
  - managed by the system
  - values:
    - private
    - submission
    - pre-evaluation, evaluation, rejected-TDG, accepted-TDG
    - pre-validation, validation, rejected-DCR Board
    - accepted

Standardization



Administrative Information Section

- The mandatory registration status
  - a designation of the status in the registration life-cycle of an administered item
  - managed by the system
  - values:
    - private
    - candidate
    - standard
    - superseded
    - deprecated



ISOcat doesn’t yet provide means to supersede or deprecate a data category

Administrative Information Section

- The effective date
  - the date a data category specification has/will become available to DCR users
- The until date
  - the date a data category specification is no longer effective in the registry; this information is set when the registration status of the data category specification changes to *deprecated* or *superseded*

ISOcat isn't acting on these dates (yet)

Description Section

- The mandatory profile
  - attribute used to relate the current data category specification to one or several thematic domains treated by ISO/TC 37 (for example, morphosyntax, syntax, metadata, language description, etc.)
  - the value of profile defaults to *Private*
  - submission for standardization requires the selection of at least one thematic domain profile because it is the relevant TDG that is responsible for maintenance of standardized data category specifications
  - if multiple profiles from multiple TDGs are selected one TDG will still be responsible, but the other TDGs will be involved in the harmonization process

When you need a new profile, contact the DCR Board or the ISOcat system administrator

Thematic Domain Groups

TDG 1: Metadata

TDG 2: Morphosyntax

TDG 3: Semantic Content Representation

TDG 4: Syntax

TDG 5: Machine Readable Dictionary

TDG 6: Language Resource Ontology

TDG 7: Lexicography

TDG 8: Language Codes

TDG 9: Terminology

TDG 11: Multilingual Information Management

TDG 12: Lexical Resources

TDG 13: Lexical Semantics

TDG 14: Source Identification

- TDGs are the owner and guardians of a coherent subset of the DCR
- TDGs own one or more profiles

- Each TDG has a chair
- A number of judges (assigned by SC P members)
- A number of expert members (up to 50%)

- TDGs are constituted at the TC37/SC plenary
- New TDGs need to be proposed by a SC
  1. Translation
  2. Sign language
  3. Audio

Data Element Name Sections

- used to record names for the data category as used in a given database, format or application
- language independent
- attributes:
  - the mandatory data element name
    - one identifier (word, multi-word unit or (alpha)numeric representation
  - the mandatory source
    - the database, format or application in which the data element name is used

ISOcat needs to better support keeping sources in sync

Data Category example

- Data category: */Grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
  - Descriptive part:
    - Data Element Name: GramGender in Text Meaning Representation
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
    - French definition: Catégorie fondée (selon la langue) sur la distinction naturelle entre les sexes ou d'autres critères formels.
  - Linguistic part:
    - Morposyntax conceptual domain: */male/, /feminine/, /neuter/*
    - French conceptual domain: */male/, /feminine/*

Working and object languages

- Working language:
  - language used to describe objects
- Object language:
  - language being described

You can describe properties of the object language French in the working language Dutch:

In de Franse taal worden vrouwelijke en mannelijk zelfstandige naamwoorden onderscheiden.

Language Section

- The English language section is mandatory and has to contain at least one name and one definition
- Additional language sections can be added as needed, and should translate at least the definition of the English language section
- When you need an additional language, contact the DCR system administration

Language Section

- The Name Section
  - records a possible name for the data category in a specific language
  - status:
    - standardized name
    - preferred name
    - admitted name
    - superseded name
    - deprecated name

Language Section

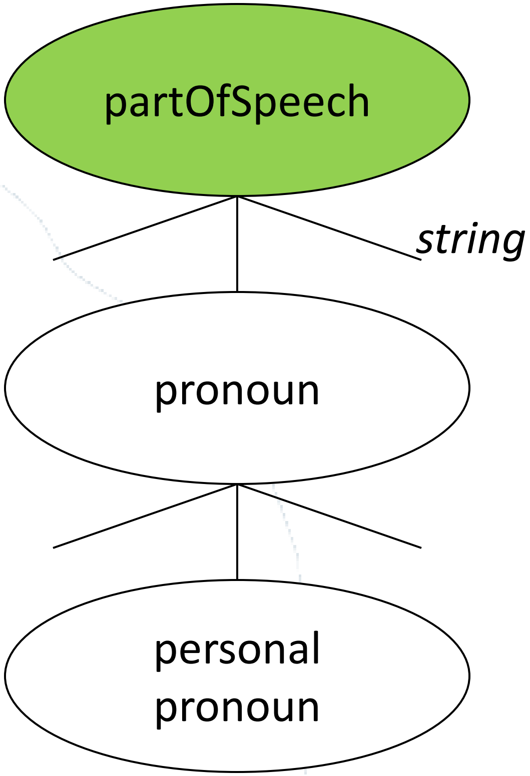
- The Definition Section
  - definition of the data category concept associated with the data category, written in the language of the language section
  - attributes:
    - definition:
      - definitive formulation that should be general enough to apply to all thematic domains and implementations of the data category
    - source:
      - from which the definition has been borrowed or adapted
    - note:
      - any additional information about the definition

ISO 704 Terminology work — principles and methods

- Intensional definitions:
  - They should consist of a single sentence fragment;
  - They should begin with a meaningful *broader concept, either immediately above or at a higher* level of the data category concept being defined;
  - They should list critical and delimiting *characteristic(s) that distinguish the concept from other* related *concepts.*
- Actual concept systems, such as are implied here by the reference to broader and related concepts, should be modeled in Relation Registries outside the DCR. Furthermore, different domains and communities of practice may differ in their choice of the immediate broader concept, depending upon any given ontological perspective. Harmonized definitions for shared DCs should attempt to choose generic references insofar as possible.

Data Category relationships

- Value domain membership
- Subsumption relationships between simple data categories (legacy)
- Relationships between complex data categories are not stored in the DCR



No ontological relationships?

- Rationale:
  - Relation types and modeling strategies for a given data category may differ from application to application;
  - Motivation to agree on relation and modeling strategies will be stronger at individual application level;
  - Integration of multiple relation structures in DCR itself could lead to endless ontological clutter.

Usage of is-a relationships between simple DCs

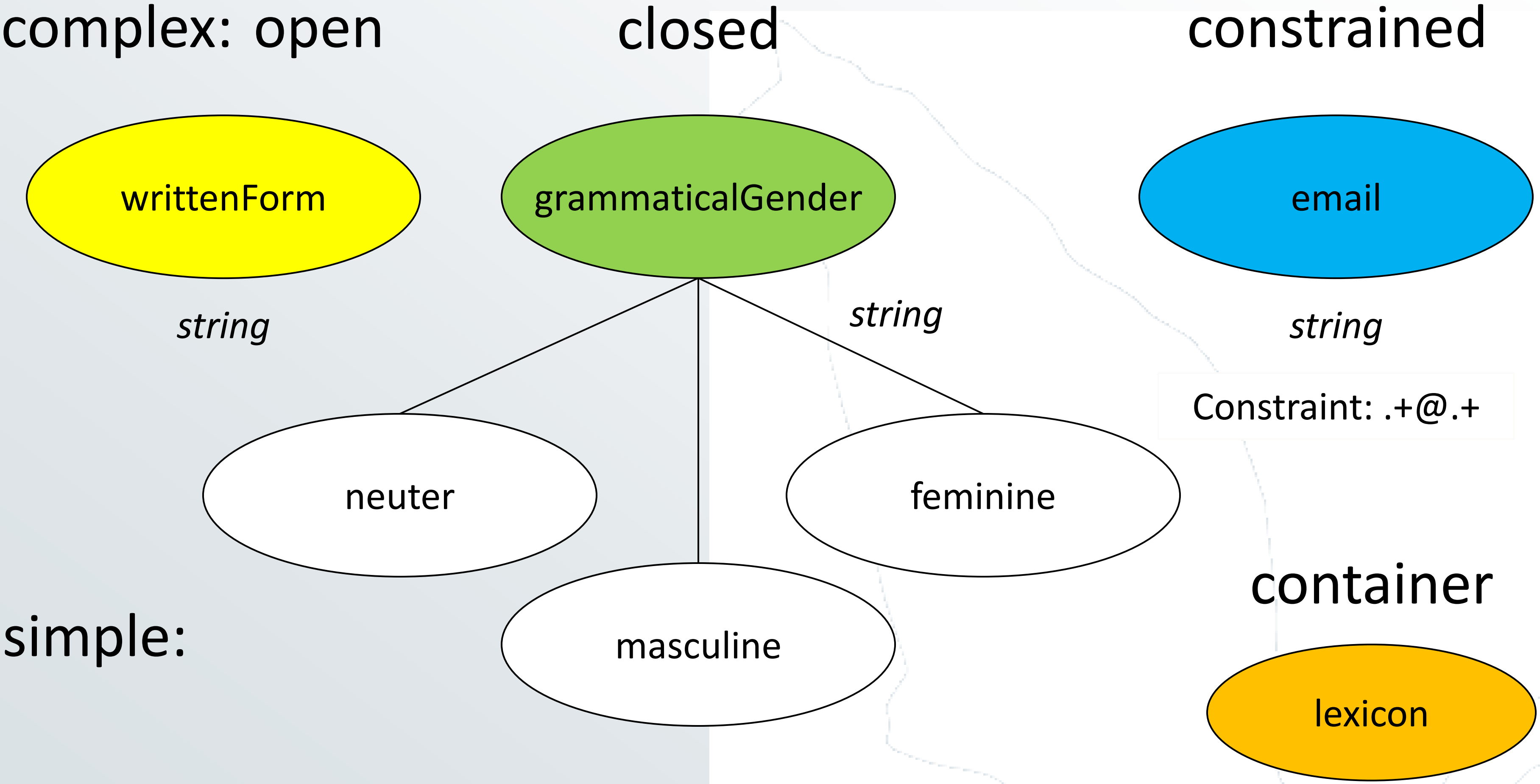
Data category	Morposyntax	Terminology
● */partOfSpeech/*	X	X
○ */adjective/*	X	X
○ */ordinalAdjective/*	X	
○ */participleAdjective/*	X	
○ */qualifierAdjective/*	X	
○ */adposition/*	X	X
○ */circumposition/*	X	
○ */preposition/*	X	
○ */postposition/*	X	

Language Section

- The examples
  - a sample instance reflecting the data category
  - should be limited to those that illustrate the data category in general, excluding language specific usage, which should be documented in a language specific Linguistic Section
  - may be accompanied by the source of the example
- The explanations
  - any additional information about the data category that would not be relevant for a definition (for example, more precise linguistic background concerning the use of the data category)
  - may be accompanied by the source from which the explanation has been borrowed or adapted
- The notes
  - any additional information associated with the data category, excluding technical information that would normally be described in an explanation

www.isocat.org

Data Category types

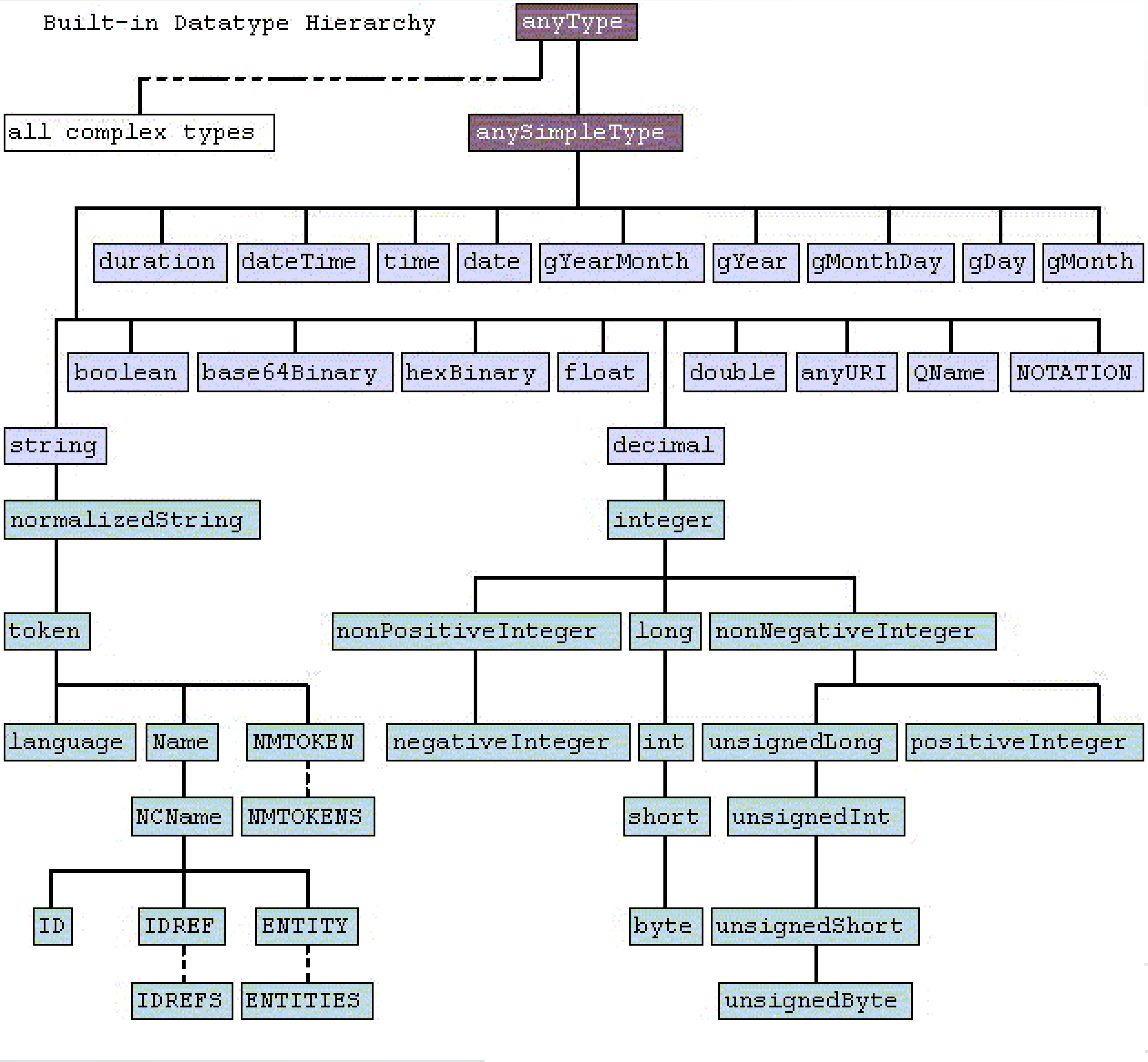


Conceptual domains

- The mandatory data type
  - – the data type, as defined for W3C XML Schema, of this complex data category
  - – the default data type is *string*

- Just the data type is enough for an open data category

W3X XML Schema data types



Constrained Conceptual Domain

- The mandatory constraint:
  - allows users to express constraints on the possible values of a conceptual domain associated with a given data type in a rule language suitable for the schema in question
  - rule languages currently ‘supported’:
    - Schematron rules
    - Object Constraint Language
    - Semantic Web Rule Language
    - Relax NG datatype parameter
    - XML Schema regular expression
    - XML Schema facet
  - the DCR doesn’t do any interpretation of the rules, this has to be done by the user/TDG/DCR Board

- ISOcat does check if its valid XML (when applicable) and includes the constraints in exports (when applicable)

Profile Value Domain

- set of permissible values for a specific profile
- each value is represented by a simple data category
- the simple data category needs to be a member of the profile

ISOcat makes an exception for the *Private* profile value domain which can contain any simple data category

- Standardized data categories can’t have a *Private* profile value domain

Usage of is-a relationships between simple DCs

Data category	Morposyntax	Terminology
● /partOfSpeech/	X	X
○ /adjective/	X	X
○ /ordinalAdjective/	X	
○ /participleAdjective/	X	
○ /qualifierAdjective/	X	
○ /adposition/	X	X
○ /circumposition/	X	
○ /preposition/	X	
○ /postposition/	X	

Linguistic Sections

- used to specify the behaviour of a complex data category in a specific object language
- Language specific examples, explanations and notes
- Refinement of the conceptual domain:
  - (additional) constraints for open and constrained complex data categories
  - subset value domains for closed complex data categories

Data Category example

- Data category: */Grammatical gender/*
  - Administrative part:
    - Identifier: grammaticalGender
    - PID: http://www.isocat.org/datcat/DC-1297
  - Descriptive part:
    - English definition: Category based on (depending on languages) the natural distinction between sex and formal criteria.
    - French definition: Catégorie fondée (selon la langue) sur la distinction naturelle entre les sexes ou d'autres critères formels.
  - Linguistic part:
    - Morposyntax conceptual domain: */male/, /feminine/, /neuter/*
    - French conceptual domain: */male/, /feminine/*

Hierarchical Simple Data Categories

- Simple data categories can be put in a subsumption (is-a) hierarchy
  - allows different levels of granularity in a value domain
  - make large value domains manageable
  - a simple data category can be only a member of one hierarchy, i.e., it can have only one parent

Usage of is-a relationships between simple DCs

Data category	Morposyntax	Terminology
● */partOfSpeech/*	X	X
○ */adjective/*	X	X
○ */ordinalAdjective/*	X	
○ */participleAdjective/*	X	
○ */qualifierAdjective/*	X	
○ */adposition/*	X	X
○ */circumposition/*	X	
○ */preposition/*	X	
○ */postposition/*	X	

Changes

- Each time you save a data category you're asked to enter a description of what you've changed
- These descriptions are available in the history log of each data category
- The DCIF contains the first (upon time of creation) and the last change description

Checker

- Not all mandatory parts of the specification has to be filled at once
- The check will tell you what is still missing
- Some errors or warnings can only be fixed by issues change requests to a standardized data category
  - Membership of another profile

Bulk import

- The ISOcat system administrator can import bulks of new Data Categories or updates
  - Provide a valid DCIF document
    - Latest version of the schema available at
      http://www.isocat.org/12620/
    - See also the DCIF export of a DCS