Controlled Vocabularies in GInAS

Thomas Balzer, 08.09.2015, Uppsala

Definitions

Controlled Vocabulary

- A controlled vocabulary (CV) is a list of words and phrases for epresenting a distinct and well defined piece of information, which can be reused multiple times in information systems
- CVs consist of terms that represent a certain fact or information
- A term is a bijective projection of a word or phrase and a code
- A term can be connected to further information like a description or additional data elements

Controlled Vocabularies vs. Natural Language

☺ Unambigous spelling

☺ Homogenous use of vocabulary

☺ Use of code is easier to handle in databases

☺ Easier translation in different languages

☺ Increases the findability of data

☺ Can deal with homonyms or synonyms

☺ Can provide background information

☹ Less flexibility to express facts

☹ Missing values have to be defined first

Controlled Vocabularies in GInAS

Use of Controlled Vocabularies in GInAS

- **GInAS maintains substance date in a structured way**
- **Simple model with domain, value and code**
- **To improve validity of data it makes use of lot of CVs**
- **GInAs uses two kinds of CVs**
  - Predefined CVs that are defined in the data model
    - Country
    - Language
    - Jurisdiction
  - Configurated CVs that are created on request
    - Substance Properties
    - CVs that are used by properties

Controlled Vocabularies in GInAS

- DOCUMENT_TYPE
- JURISDICTION
- NAME_TYPE
- OPTICAL_ACTIVITY
- STEREOCHEMISTRY_TYPE
- PROPERTY_NAME
- PROPERTY_TYPE
- PROTEIN_TYPE
- AMOUNT_UNIT
- AMOUNT_TYPE
- …

Substance Class

- Chemical
- Protein
- Nucleic Acid
- Polymer
- Structurally Diverse
- Mixture

Language

- **ISO 639-1:2002 – Codes for the representation of names of languages – Part 1: Alpha-2 code**
- **ISO 639-2:1998 – Codes for the representation of names of languages – Part 2: Alpha-3 code**

- **Example:**

Language name:	German
Native name:	Deutsch
639-1:	de
639-2/T:	deu
639-2/B:	ger

T = terminological code, derived from the native name
B = bibliographic code, derived from the English name

Amount Unit

- **Unit of Measurement**
- **ISO 11240:2012**
- **Needs additional Properties**
  - UCUM coding for exchange
  - Classification of Units
    - Mass
    - Amount
    - Time
    - …
- **Other Sources**

Solubility

- **slightly soluble**
- **soluble**
- **Insoluble**

- **Parameter:**
  - Solvent
  - Temperature
  - Pressure?

- **When to use which value?**
- **Are there agreed upper or lower boundaries for the amount?**

Requirements for Controlled Vocabularies

Functional Requirements

- Clear Definitions
- Multilingual Terms
- Different Jurisdictions
- Synonyms
- Historical Terms / Versions
- Hierarchies
- Replacement Terms
- References to External Sources
- Alternative Code Systems
- Universal Usable Code System
- Additional Attributes can be defined

Requirement for a Maintenance System

- **Keeps track on all changes**
- **Keeps versions of all previous versions**
- **Allows fast track modifications**
- **Provides a workflow to maintain change requests**
  - New Terms
  - Change requests for Terms
  - Translation and documentation
  - Quality management
- **Provides services to make controlled vocabularies publicly available**

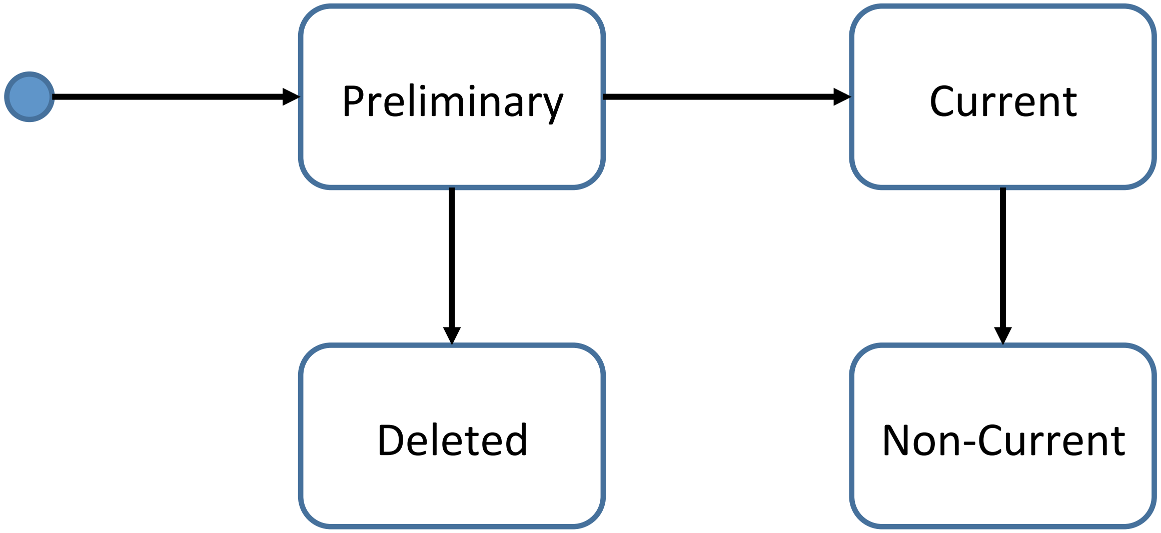
Publication of Controlled Vocabularies

- **REST service**
- **SOAP service**
  - Common Terminology Services CTS2 defined by Object Management Group
- **A set of services has to be provided**
  - Catalog Service
  - Description Service
  - Mapping Service

Versioning of Terms

- Only the current version of a term has a functional id
- Only the current version of a term has a status CURRENT
- Each version of a term has ist own term id
- Each version keeps a reference to the previous version
- A revision counter is increased every time a term is modified
- A modified term gets the status NON-CURRENT and the functional id is set to null

Status Model

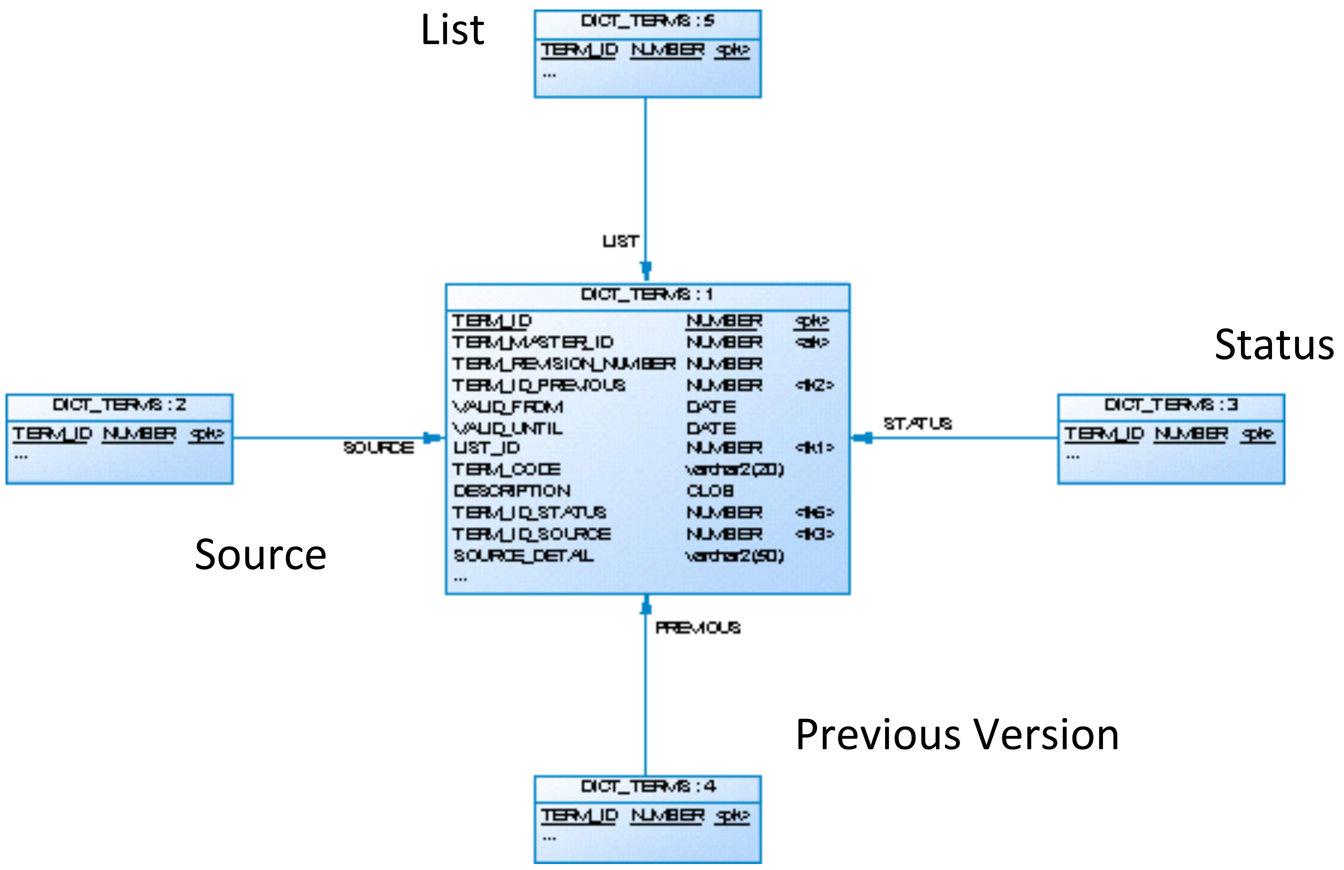


Integration of Controlled Vocabularies in GInAS

- GInAS makes a reference to the term id
  ➔ Modifying a term does not change the used term
- New term are used only upon request
- All terms seem to be in a simple lookup table

Datamodel

Terms



List

DICT_TERMS : 5

TERM_ID	NUMBER	<pk>
...		

DICT_TERMS : 1

TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
TERM_ID_STATUS	NUMBER	<fk6>
TERM_ID_SOURCE	NUMBER	<fk3>
SOURCE_DETAIL	varchar2(50)	
...		

Status

DICT_TERMS : 3

TERM_ID	NUMBER	<pk>
...		

Source

DICT_TERMS : 2

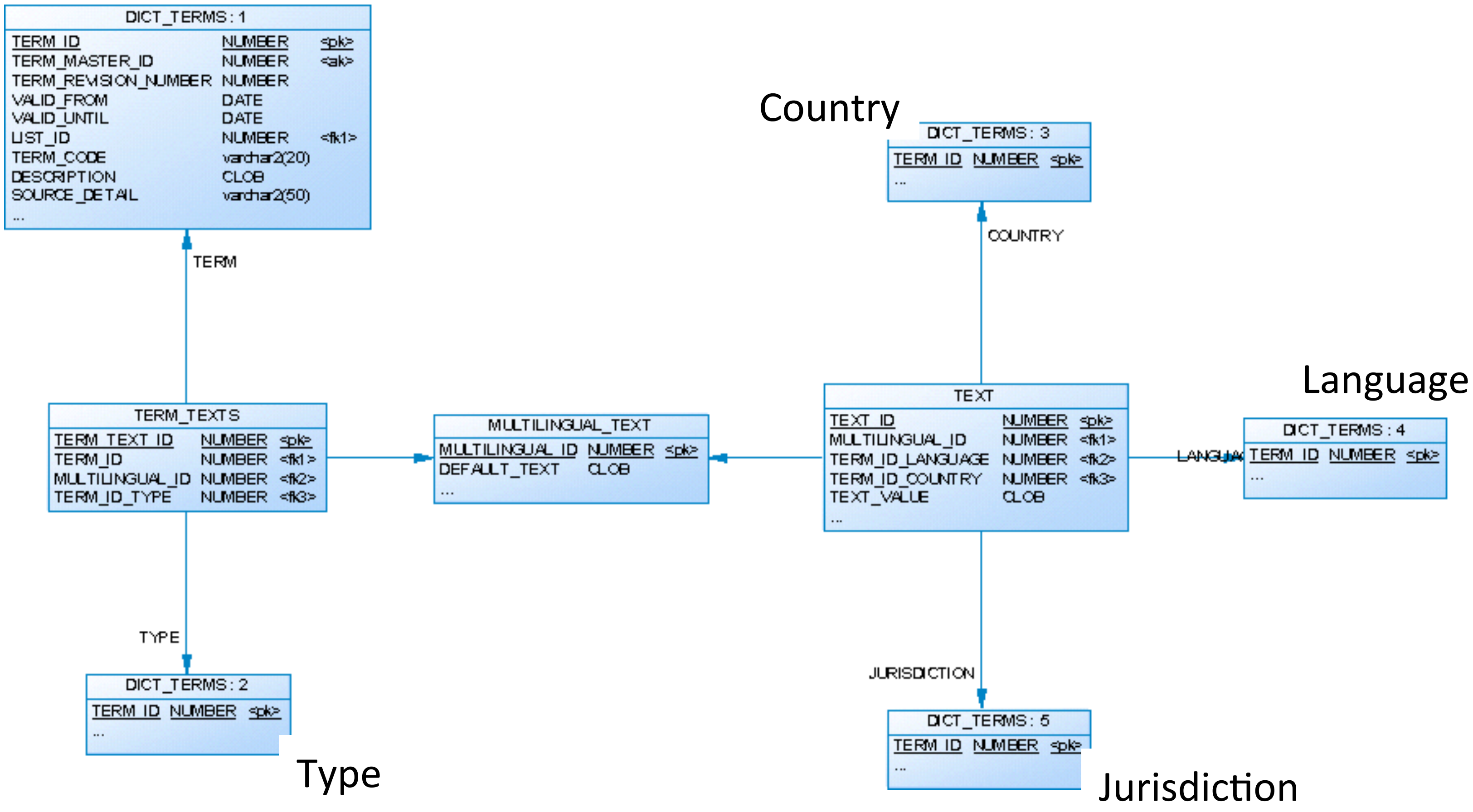
TERM_ID	NUMBER	<pk>
...		

Previous Version

DICT_TERMS : 4

TERM_ID	NUMBER	<pk>
...		

Multilingual Texts



DICT_TERMS : 1

TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
SOURCE_DETAIL	varchar2(50)	
...		

TERM_TEXTS

TERM_TEXT_ID	NUMBER	<pk>
TERM_ID	NUMBER	<fk1>
MULTILINGUAL_ID	NUMBER	<fk2>
TERM_ID_TYPE	NUMBER	<fk3>

MULTILINGUAL_TEXT

MULTILINGUAL_ID	NUMBER	<pk>
DEFAULT_TEXT	CLOB	
...		

TEXT

TEXT_ID	NUMBER	<pk>
MULTILINGUAL_ID	NUMBER	<fk1>
TERM_ID_LANGUAGE	NUMBER	<fk2>
TERM_ID_COUNTRY	NUMBER	<fk3>
TEXT_VALUE	CLOB	
...		

DICT_TERMS : 2

TERM_ID	NUMBER	<pk>
...		

DICT_TERMS : 3

TERM_ID	NUMBER	<pk>
...		

DICT_TERMS : 4

TERM_ID	NUMBER	<pk>
...		

DICT_TERMS : 5

TERM_ID	NUMBER	<pk>
...		

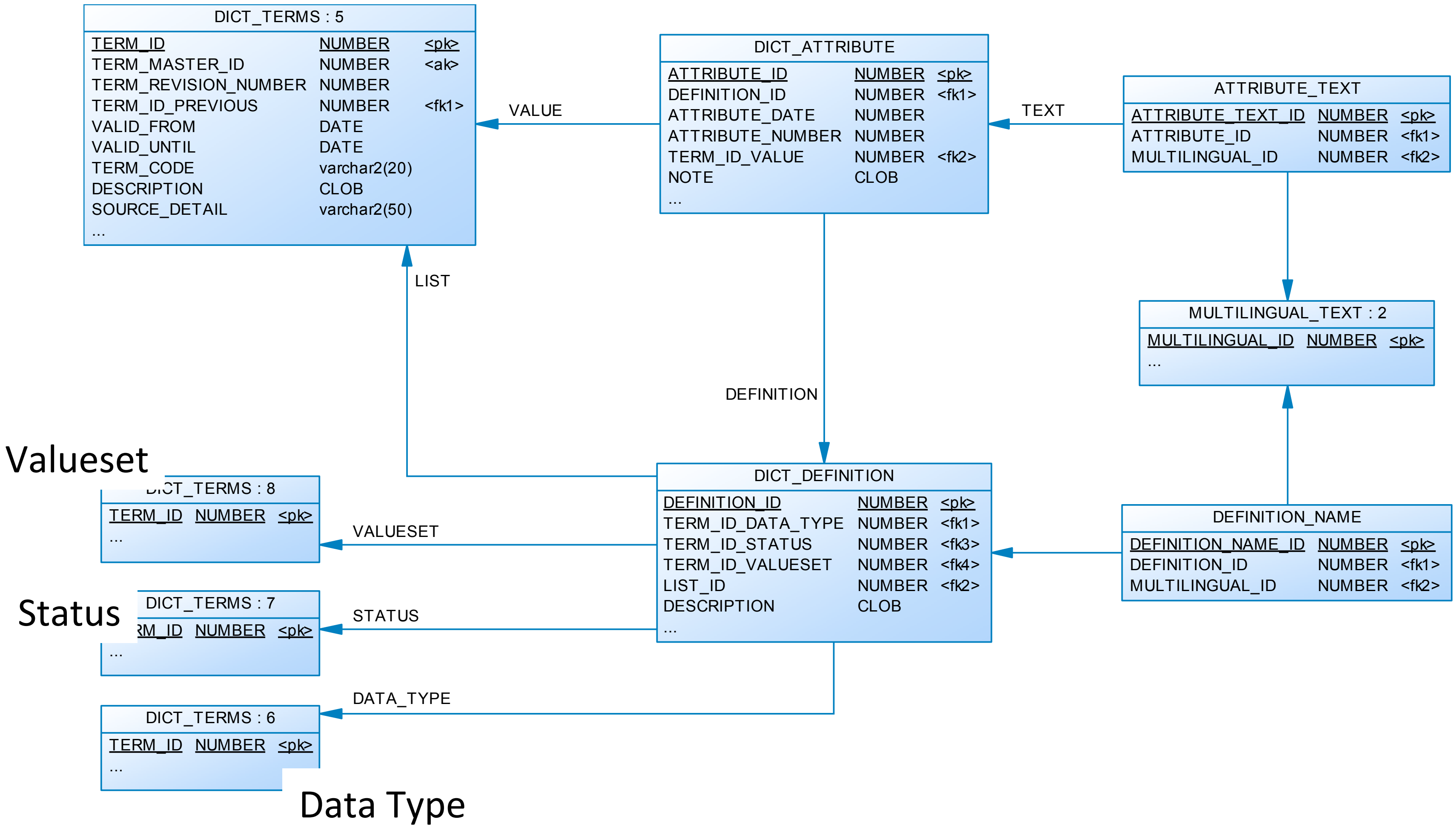
Country

Language

Type

Jurisdiction

Additional Attributes



DICT_TERMS : 5		
TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk1>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
SOURCE_DETAIL	varchar2(50)	
...		

DICT_ATTRIBUTE		
ATTRIBUTE_ID	NUMBER	<pk>
DEFINITION_ID	NUMBER	<fk1>
ATTRIBUTE_DATE	NUMBER	
ATTRIBUTE_NUMBER	NUMBER	
TERM_ID_VALUE	NUMBER	<fk2>
NOTE	CLOB	
...		

ATTRIBUTE_TEXT		
ATTRIBUTE_TEXT_ID	NUMBER	<pk>
ATTRIBUTE_ID	NUMBER	<fk1>
MULTILINGUAL_ID	NUMBER	<fk2>

MULTILINGUAL_TEXT : 2		
MULTILINGUAL_ID	NUMBER	<pk>
...		

DICT_DEFINITION		
DEFINITION_ID	NUMBER	<pk>
TERM_ID_DATA_TYPE	NUMBER	<fk1>
TERM_ID_STATUS	NUMBER	<fk3>
TERM_ID_VALUESET	NUMBER	<fk4>
LIST_ID	NUMBER	<fk2>
DESCRIPTION	CLOB	
...		

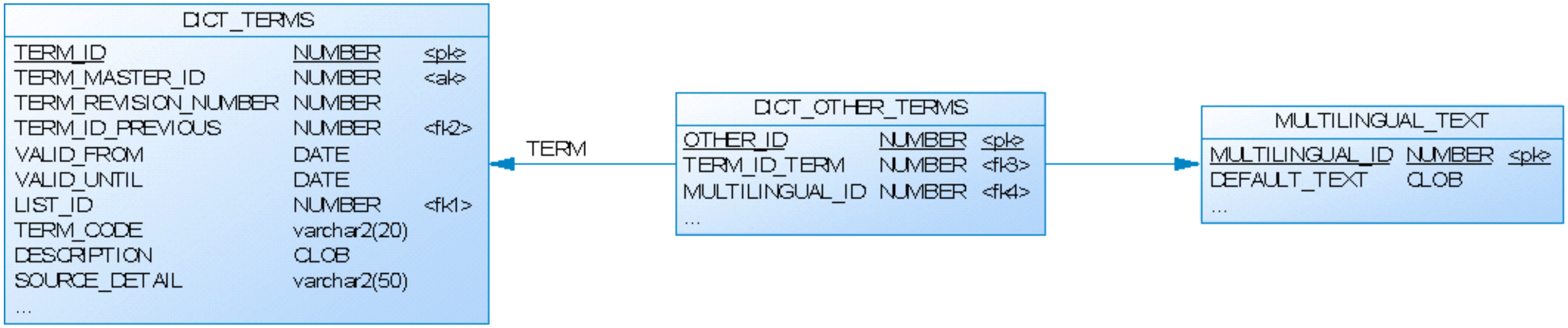
DEFINITION_NAME		
DEFINITION_NAME_ID	NUMBER	<pk>
DEFINITION_ID	NUMBER	<fk1>
MULTILINGUAL_ID	NUMBER	<fk2>

DICT_TERMS : 8		
TERM_ID	NUMBER	<pk>
...		

DICT_TERMS : 7		
TERM_ID	NUMBER	<pk>
...		

DICT_TERMS : 6		
TERM_ID	NUMBER	<pk>
...		

Synonyms



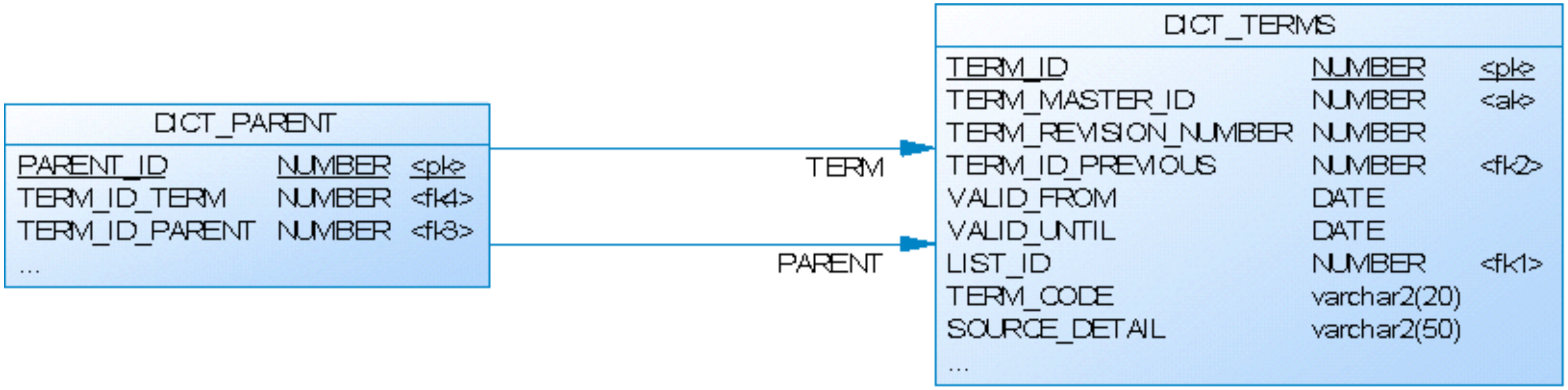
DICT_TERMS		
TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
SOURCE_DETAIL	varchar2(50)	
...		

TERM

DICT_OTHER_TERMS		
OTHER_ID	NUMBER	<pk>
TERM_ID_TERM	NUMBER	<fk3>
MULTILINGUAL_ID	NUMBER	<fk4>
...		

MULTILINGUAL_TEXT		
MULTILINGUAL_ID	NUMBER	<pk>
DEFAULT_TEXT	CLOB	
...		

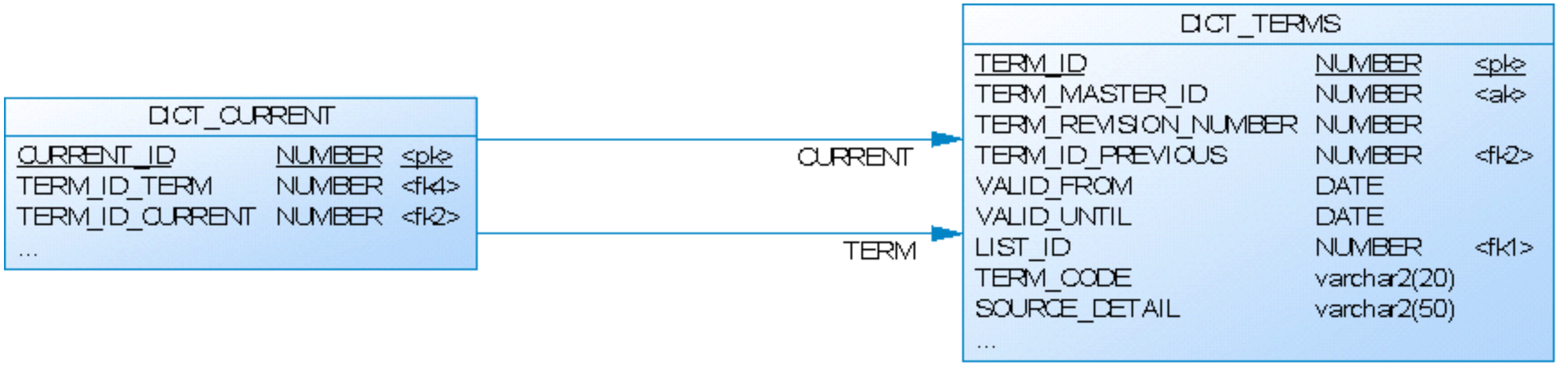
Hierarchy



DICT_PARENT		
PARENT_ID	NUMBER	<pk>
TERM_ID_TERM	NUMBER	<fk4>
TERM_ID_PARENT	NUMBER	<fk3>
…		

DICT_TERMS		
TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
SOURCE_DETAIL	varchar2(50)	
…		

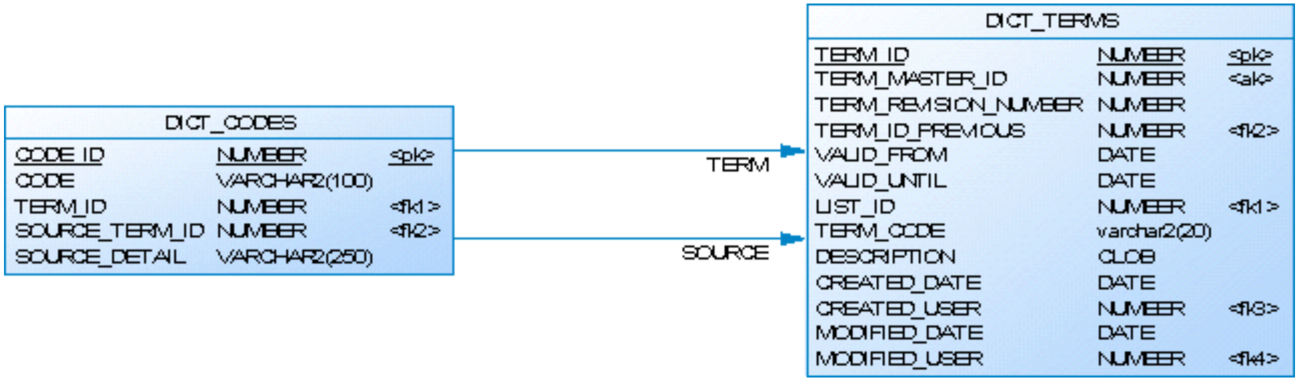
Replacement Terms



DICT_CURRENT		
CURRENT_ID	NUMBER	<pk>
TERM_ID_TERM	NUMBER	<fk4>
TERM_ID_CURRENT	NUMBER	<fk2>
...		

DICT_TERMS		
TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
SOURCE_DETAIL	varchar2(50)	
...		

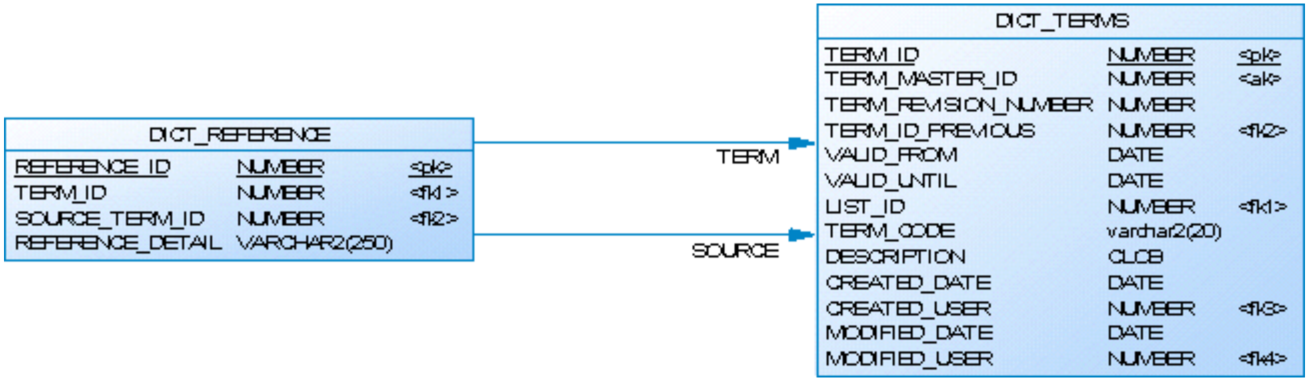
Alternative Codes



DICT_CODES		
CODE_ID	NUMBER	<pk>
CODE	VARCHAR2(100)	
TERM_ID	NUMBER	<fk1>
SOURCE_TERM_ID	NUMBER	<fk2>
SOURCE_DETAIL	VARCHAR2(250)	

DICT_TERMS		
TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
CREATED_DATE	DATE	
CREATED_USER	NUMBER	<fk3>
MODIFIED_DATE	DATE	
MODIFIED_USER	NUMBER	<fk4>

References



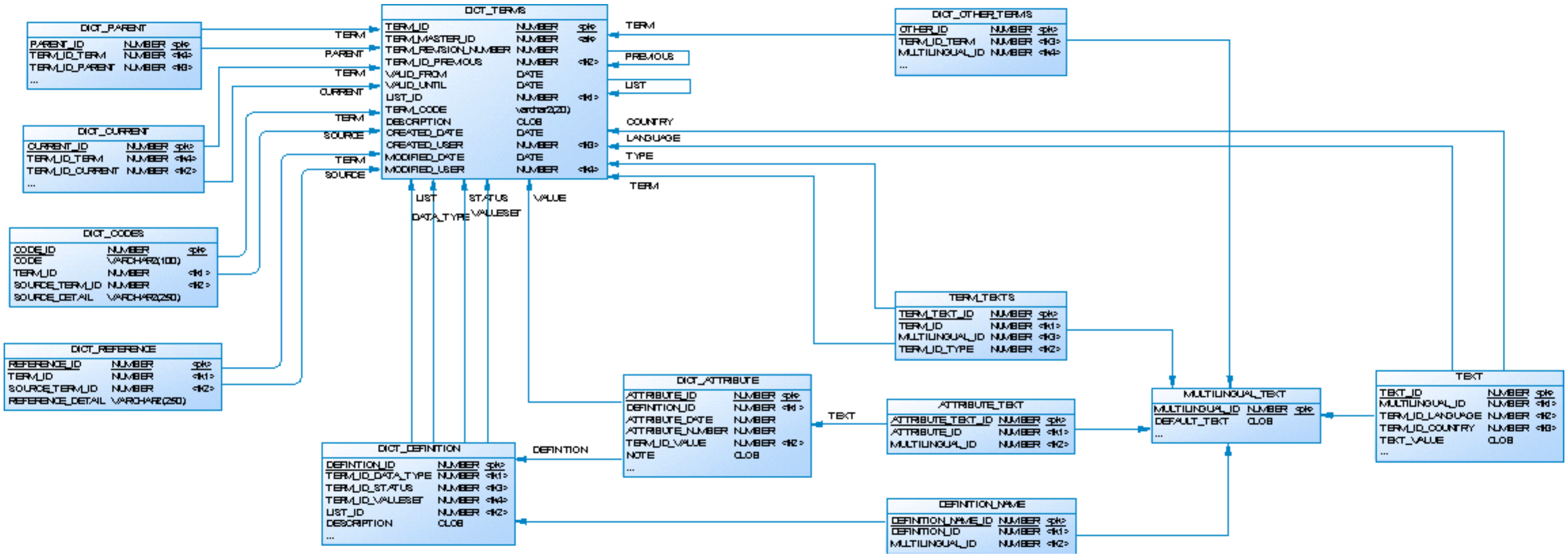
DICT_REFERENCE

REFERENCE_ID	NUMBER	<pk>
TERM_ID	NUMBER	<fk1>
SOURCE_TERM_ID	NUMBER	<fk2>
REFERENCE_DETAIL	VARCHAR2(250)	

DICT_TERMS

TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
CREATED_DATE	DATE	
CREATED_USER	NUMBER	<fk3>
MODIFIED_DATE	DATE	
MODIFIED_USER	NUMBER	<fk4>

Complete Datamodel



DICT_PARENT

PARENT_ID	NUMBER	<pk>
TERM_ID_TERM	NUMBER	<fk4>
TERM_ID_PARENT	NUMBER	<fk3>
...		

DICT_CURRENT

CURRENT_ID	NUMBER	<pk>
TERM_ID_TERM	NUMBER	<fk4>
TERM_ID_CURRENT	NUMBER	<fk2>
...		

DICT_CODES

CODE_ID	NUMBER	<pk>
CODE	VARCHAR2(100)	
TERM_ID	NUMBER	<fk1>
SOURCE_TERM_ID	NUMBER	<fk2>
SOURCE_DETAIL	VARCHAR2(250)	

DICT_REFERENCE

REFERENCE_ID	NUMBER	<pk>
TERM_ID	NUMBER	<fk1>
SOURCE_TERM_ID	NUMBER	<fk2>
REFERENCE_DETAIL	VARCHAR2(250)	

DICT_TERMS

TERM_ID	NUMBER	<pk>
TERM_MASTER_ID	NUMBER	<ak>
TERM_REVISION_NUMBER	NUMBER	
TERM_ID_PREVIOUS	NUMBER	<fk2>
VALID_FROM	DATE	
VALID_UNTIL	DATE	
LIST_ID	NUMBER	<fk1>
TERM_CODE	varchar2(20)	
DESCRIPTION	CLOB	
CREATED_DATE	DATE	
CREATED_USER	NUMBER	<fk3>
MODIFIED_DATE	DATE	
MODIFIED_USER	NUMBER	<fk4>

DICT_OTHER_TERMS

OTHER_ID	NUMBER	<pk>
TERM_ID_TERM	NUMBER	<fk3>
MULTILINGUAL_ID	NUMBER	<fk4>
...		

TERM_TEXTS

TERM_TEXT_ID	NUMBER	<pk>
TERM_ID	NUMBER	<fk1>
MULTILINGUAL_ID	NUMBER	<fk3>
TERM_ID_TYPE	NUMBER	<fk2>

DICT_ATTRIBUTE

ATTRIBUTE_ID	NUMBER	<pk>
DEFINITION_ID	NUMBER	<fk1>
ATTRIBUTE_DATE	NUMBER	
ATTRIBUTE_NUMBER	NUMBER	
TERM_ID_VALUE	NUMBER	<fk2>
NOTE	CLOB	
...		

ATTRIBUTE_TEXT

ATTRIBUTE_TEXT_ID	NUMBER	<pk>
ATTRIBUTE_ID	NUMBER	<fk1>
MULTILINGUAL_ID	NUMBER	<fk2>

MULTILINGUAL_TEXT

MULTILINGUAL_ID	NUMBER	<pk>
DEFAULT_TEXT	CLOB	
...		

TEXT

TEXT_ID	NUMBER	<pk>
MULTILINGUAL_ID	NUMBER	<fk1>
TERM_ID_LANGUAGE	NUMBER	<fk2>
TERM_ID_COUNTRY	NUMBER	<fk3>
TEXT_VALUE	CLOB	
...		

DICT_DEFINITION

DEFINTION_ID	NUMBER	<pk>
TERM_ID_DATA_TYPE	NUMBER	<fk1>
TERM_ID_STATUS	NUMBER	<fk3>
TERM_ID_VALUESET	NUMBER	<fk4>
LIST_ID	NUMBER	<fk2>
DESCRIPTION	CLOB	
...		

DEFINITION_NAME

DEFINITION_NAME_ID	NUMBER	<pk>
DEFINITION_ID	NUMBER	<fk1>
MULTILINGUAL_ID	NUMBER	<fk2>

Contact Information

Thomas Balzer (BfArM)

Mail: thomas.balzer@bfarm.de

Thank you for your attention!