Semantic Interoperability Courses

Course Module 3
Reference Data Management

V0.10
June 2014
ISA Programme, Action 1.1

Disclaimer

This training material was prepared for the ISA programme of the European Commission by PwC EU Services.

The views expressed in this report are purely those of the authors and may not, in any circumstances, be interpreted as stating an official position of the European Commission.

The European Commission does not guarantee the accuracy of the information included in this study, nor does it accept any responsibility for any use thereof.

Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Commission.

All care has been taken by the author to ensure that s/he has obtained, where necessary, permission to use any parts of manuscripts including illustrations, maps, and graphs, on which intellectual property rights already exist from the titular holder(s) of such rights or from her/his or their legal representative.

Learning Objectives

By the end of this training you should have an understanding of:

- *What reference data is, its context and purpose and how it creates value for organisations.*
- *Why it is important to manage and govern the reference data lifecycle.*
- *How to work with reference data using open-source tools.*

Outline

1. Introduction: what is reference data?

- Definitions
  - Reference data
  - Reference data has many names: code list, taxonomy, thesaurus, mapping, name authority list
- Importance & relevance

2. Why must reference data be properly managed?

- What is reference data management
- Why is managing reference data important?
- Design
- Change management
- Documentation
- Harmonisation

Definitions

What is reference data?

*Reference data is small, discrete **sets of values** that are not updated as part of business transactions but are usually used to impose consistent **classification**. Reference data normally has a low update frequency. Reference data is relevant across more than one business system belonging to different organisations and sectors.*

European Commission – ISA Programme, 2014 (1)

Example: Country Code

Named Authority Lists

The table below displays an extract of the "Countries" code list as published on the Metadata Registry (MDR) of the EU:

Authority Code	Short Name	Long Name
AND	Andorra	Principality of Andorra
ALB	Albania	Republic of Albania
AUT	Austria	Republic of Austria
BIG	Bosnia and Herzegovina	Bosnia and Herzegovina
…	…	…

http://publications.europa.eu/mdr/resource/authority/country/html/countries-eng.html#description

Reference data has many names

What is considered reference data?

- **Code list:** Complete set of data element values of a coded simple data element [ISO 9735-1:2002, 4.14]
- **Taxonomy:** scheme of categories and subcategories that can be used to sort and otherwise organize items of knowledge or information [ISO/DIS 25964-2].
- **Thesaurus:** controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms [ISO 25964-1:2011]
- **Mapping:** relationship between a concept in one vocabulary and one or more concepts in another [ISO/DIS 25964-2].
- **Name authority list:** controlled vocabulary for use in naming particular entities consistently [ISO/DIS 25964-2]

Source: https://joinup.ec.europa.eu/svn/adms/ADMS_v1.00/ADMS_SKOS_v1.00.html

Importance

Where is reference data used?

Within information systems

- For categorising and identifying data
- E.g. assigning personnel to a department from a list of predefined values

Between Information Systems

- For information sharing
- E.g. using a code list to describing the context of data which is exchanged between systems over different member states. This ensures that member states are 'talking' about the same data.

Reference data… is just data!

Relevance of *common* reference data

Why is common reference data important?

To avoid semantic interoperability conflicts

- By using a common set of values for describing data which is exchanged between different systems, interoperability conflicts can be avoided.
- Please refer to training module 1 for more information on interoperability concepts and more specifically semantic interoperability

To avoid the need for mappings

- Mappings between different value sets of reference data are often inaccurate. By using common value sets of reference data across domains and IT systems, the need for creating mappings can be avoided.

Outline

1. Introduction: what is reference data?

- Definitions
  - Reference data
  - Reference data has many names: code list, taxonomy, thesaurus, mapping, name authority list
- Importance & relevance

2. Why must reference data be properly managed?

- What is reference data management
- Why is managing reference data important?
- Design
- Change management
- Documentation
- Harmonisation

Definitions

What is reference data management?

Reference data management comprises planning, implementation & control activities to **ensure consistency** with "golden version" of contextual data values.

Reference Data Management is control over defined domain values (also known as vocabularies), including control over standardized terms, code values and other unique identifiers, business definitions for each value, business relationships within and across domain value lists, and the consistent, shared used of accurate, timely and relevant reference data values to classify and categorize data

DAMA International, 2009,
http://www.dama.org/i4a/pages/index.cfm?pageid=1

Relevance

Why is metadata management important?

- To ensure the use of a common setting
- To ensure continuity and quality of service
- To take decisions and manage changes in a controlled fashion
- To prevents conflicts between versions (version control)
- To improve data quality

Reference data management

Lifecycle

1. Data Design

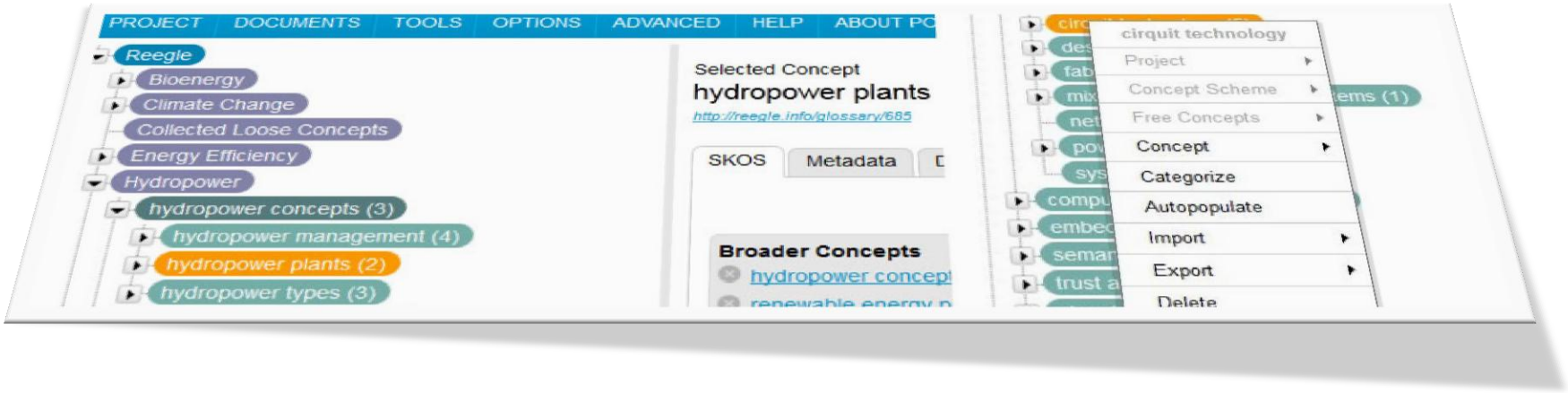
What	Why	How

• Develop thesauri, value sets, code lists, etc.
• Select and reuse existing reference data sets
	• Impose consistent classification of data
	• Improve data quality
	• Reduce interoperability issues
		• **Tools**
		  • VocBench
		  • PoolParty
		  • ListPoint

1. Reference Data Design | Tools

PoolParty

PoolParty is a tool for creating thesauri, taxonomies and knowledge graphs based on W3C standards such as SKOS, RDF and SPARQL.
(Semantic Web Company, 2014)



1. Reference Data Design | Tools

VocBench

VocBench is a web-based, multilingual, vocabulary editing and workflow tool (W3C, 2001). It manages thesauri , authority lists and glossaries using SKOS-XL. (FAO, 2014)

Listpoint

Listpoint is an open reference data platform combined with online tools to find and combine data standards and code lists. Moreover, it helps users to make datasets interoperable and kept up-to-date with updates. (Listpoint, 2014).

2. Change Management

What	Why	How
• A combination of management processes for incorporating changes to value sets.	• Maintaining control over the value sets and the change process • Taking into account the needs of stakeholders when adapting reference data	• Defining each step in the change process and assigning roles which are described in a governance structure • Incorporating quality control measures

2. Change Management

In a managed master data environment, specific individuals have the role of a business data steward. They have the authority to create, update, and retire reference data values, and to a lesser extent, in some circumstances, master data values,. Business data stewards work with data professionals to ensure the highest quality reference and master data. Many organizations define more specific roles and responsibilities, with individuals often performing more than one role.

Steps in change management are:

1. Create and receive change requests
2. Identify the related stakeholders and understand their interest
3. Identify and evaluate the impacts of the proposed changes
4. Decide to accept or reject the change, or recommend a decision to management or governance
5. Review and approve or deny the recommendation, if needed
6. Communicate the decision to stakeholders prior to making the change
7. Update the data
8. Inform stakeholders the change has been made

3. Documentation

What	Why	How

• Representing the reference data value sets following international standards
• Describing the value sets in a uniform way
	• To avoid misinterpretation of the value set
	• To ensure machine-readability
	• Description: to facilitate searching and retrieving reference data from a repository
		• **Representation**
		  • SKOS
		  • GeneriCode
		  • XSD
		  • HTML
		• **Publication**
		  • Metadata Registry
		• **Description**
		  • ADMS

3. Documentation | XML representation

XML Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). It permits to represent reference data in many different ways.

XML extract for representing Andalusia in the Countries NAL

```
<record adm.status="current" date.creation="2010-01-01" IMMC.approval.date="2012-06-27"
IMMC.proposal.date="2011-10-06" pub="false" celex="false" deprecated="false" id="COU0001">

            <code-3166-1-alpha-2>AD</code-3166-1-alpha-2>
            <code-3166-1-alpha-3>AND</code-3166-1-alpha-3>
            <code-3166-1-num>020</code-3166-1-num>
            <authority-code>AND</authority-code>
            <code-iana>.ad</code-iana>
            <code-tir>AND</code-tir>
            <name><original.name>
                        <lg.version lg="cat">Andorra</lg.version>
            </original.name></name>

<label>
            <lg.version lg="bel" script="Cyrillic">Андора</lg.version>
            <lg.version lg="bos">Andora</lg.version>
            …
</label>

</record>
```

3. Documentation | SKOS representation

SKOS is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS).

SKOS extract for representing Andalusia in the Countries NAL

```
<skos:Concept rdf:about="http://publications.europa.eu/resource/authority/country/AND"
at:deprecated="false">
            <skos:inScheme
            rdf:resource="http://publications.europa.eu/resource/authority/country"/>
            <at:authority-code>AND</at:authority-code>
            <at:op-code>AND</at:op-code>
            <atold:op-code>AND</atold:op-code>
            <dc:identifier>AND</dc:identifier>
            <at:start.use>1950-05-09</at:start.use>
            <skos:prefLabel xml:lang="be">Андора</skos:prefLabel>
            <skos:prefLabel xml:lang="bs">Andora</skos:prefLabel>
            <skos:prefLabel xml:lang="bg">Андора</skos:prefLabel>
            …
</skos:Concept>
```

3. Documentation | XSD representation

XSD: XML Schemas express shared vocabularies and allow machines to carry out rules made by people.

XSD extract for representing a country in the Countries NAL

```
<!--RECORD DEFINITION-->
<xs:element name="record">
            <xs:complexType>
                        <xs:sequence>
                                    <xs:element ref="code-3166-1-alpha-2"/>
                                    <xs:element ref="code-3166-1-alpha-3"/>
                                    <xs:element ref="code-3166-1-num"/>
                                    <xs:element maxOccurs="unbounded" ref="code-3166-3"
                        minOccurs="0"/>
                                    <xs:element ref="authority-code"/>
                                    <xs:element ref="op-styleguide" minOccurs="0"/>
                                    <xs:element maxOccurs="unbounded" ref="code-iana"
                                    minOccurs="0"/>
                                    <xs:element ref="code-tir" minOccurs="0"/>
                                    <xs:element ref="name"/>
                                    <xs:element ref="label"/>
                                    …
                        </xs:sequence>
            </xs:complexType>
</xs:element >
```

3. Documentation | Genericode representation

GeneriCode

Genericode defines a standard format for defining code lists (also known as enumerations or controlled vocabularies). It contains:

- a standard model and XML representation for the **contents** of a code list;
- a standard model and XML representation for **data associated** with items in a code list;
- a standard model and XML representation for how new code lists are **derived** from existing code lists.

"Genericode not only provides a representation of the items in a code list, it also provides an audit trail for how that code list is related to previous versions of the code list, or to other code lists. This simplifies the effort of understanding how a new code list version differs from the previous version, and simplifies the effort in calculating the impact of the change on existing systems and processes." (Genericode, 2014)

3. Documentation | HTML representation

HTML (Hyper Text Markup Language) is used to describe documents

TABLE-ID: country

Current entries date: 2013-11-19

Select language: bul | spa | ces | dan | deu | est | ell | eng | fra | gle | hrv | ita | lav | lit | hun | mlt | nld | pol | por | ron | slk | slv | fin | swe

Description | Metadata codes and names | Country codes comparison and other attributes | Facts

- Authority code is the ISO 3166-1/α-3, with exceptional alpa-numerical codes when ISO code doesn't exist.
- The Authority code identifies the record that contains the short and long form of the name and other information.
- Date of event is indicated when a historical relationship occurred.

Authority code	ISO 3166-1/α-3	Short Name	Long Name	Date Of Event	Predecessor	Successor	Related To	Comments
AND	AND	Andorra	Principality of Andorra	1950 -				
ALB	ALB	Albania	Republic of Albania	1912 -			Soviet Union (SU)	Date: 2011-02-14: Comment: Satellite state of the Soviet Union (1944-1960) in the Warsaw Pact. Government extant until 1992. (Source: Wikipedia)
AUT	AUT	Austria	Republic of Austria	1950 -				
BIH	BIH	Bosnia and Herzegovina	Bosnia and Herzegovina	1992 -	Yugoslavia (YU)			Date: 2011-02-14: Comment: Opening event: Splitting of the "Socialist Federal Republic of Yugoslavia" (COU0060) into Republic of Slovenia (COU0051), Republic of Croatia

3. Documentation | ADMS description

ADMS

The Asset Description Metadata Schema (ADMS) is a common way to describe semantic interoperability assets making it possible for everyone to search and discover them.

ADMS allows public administrations, businesses, standardisation bodies and academia to (European Commission – ISA Programme, 2011):

- *"describe semantic assets in a common way so that they can be seamlessly cross-queried and discovered by ICT developers from a single access point, such as Joinup;*
- *search, identify, retrieve, compare semantic assets to be reused avoiding duplication and expensive design work through a single point of access;*
- *keep their own system for documenting and storing semantic assets;*
- *improve indexing and visibility of their own assets;*
- *link semantic assets to one another in cross-border and cross-sector settings."*

https://joinup.ec.europa.eu/asset/adms/description

3. Documentation | Publication

Metadata Registry

A best practice in reference data management is to publish value sets on an authoritative source. An example of such a source is the Metadata Registry (MDR) of the EU, which is maintained by the Publications Office. The MDR registers and maintains metadata used by European Institutions involved in the legal decision making process.

http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35343

4. Harmonisation

What	Why	How
• The alignment of structural metadata used for information exchange either through the creation of mappings between terms of two or more specifications for structural metadata or by forging a wide consensus on the use of a common specification.	• To foster interoperability with reference data value sets which are represented using a different standard	• Reference Data Mappings • Tool: Silk Workbench

4. Harmonisation | Example

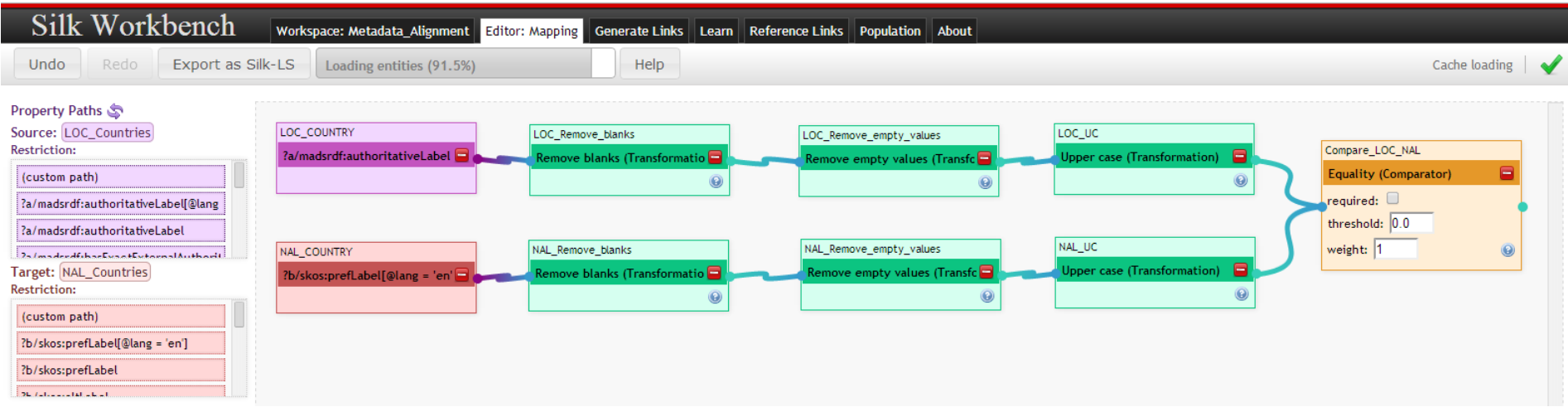
The table below shows a mapping of the Publications Office Named Authority List for countries with the ISO 3166 standard.

Authority Code	ISO 3166	Short Name	Long Name
AND	AND	Andorra	Principality of Andorra
ALB	ALB	Albania	Republic of Albania
AUT	AUT	Austria	Republic of Austria
BIH	BIH	Bosnia and Herzegovina	Bosnia and Herzegovina
…	…	…	…

http://publications.europa.eu/mdr/resource/authority/country/html/countries-eng.html#description

4. Harmonisation | Silk Workbench

Tool: [Silk Workbench]
The Silk framework is a tool for discovering relationships between data items within different Linked Data sources.



5. Implementation

What	Why	How
• Propagating reference data changes into the software development lifecycle • Manage and support the exchange of information between systems	• Coordinated use of reference data • Reference data has a lifecycle and needs to be updated • Improving reusability	• Manual or automatic propagation • In case of automatic propagation, changes to reference data into operational systems should be controlled by governance processes

Implementation of reference data in information systems

- To manage and support the exchange of information between systems, the propagation of changes to reference data is needed
- Can be done automatically or manual
- Propagation of reference data changes needs to be **part of the software development lifecycle** in order to ensure coordinated and timely updates of reference data in all information systems involved.

References

European Commission – ISA Programme. (2011). *Asset Description Metadata Schema (ADMS).* Brussels.

European Commission - ISA Programme. (2012). *D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC.* Brussels.

European Commission - ISA Programme. (2012). *Asset Description Metadata Schema for Software.* Brussels.

European Commission – ISA Programme. (2014). *D4.1. Metadata management requirements and existing solutions in EU Institutions and Member States.* Brussels.

European Commission – ISA Programme. (2014). *D4.5. Metadata alignment pilot in the EU institutions and MSs.* Brussels.

References

W3C. (2001). *VocBench.* Available at http://www.w3.org/2001/sw/wiki/VocBench.

DAMA International. (2009). *DAMA International*. Available at http://www.dama.org/

FAO. (2014). *VocBench.* Available at http://aims.fao.org/tools/vocbench-2.

Genericode. (2014). *What is 'genericode'?* Available at http://www.genericode.org/.

Listpoint. (2014). *Welcome to Listpoint.* Available at https://www.listpoint.co.uk/.

Mosley, M., Brackett, M., Earley, S., & Henderson, D. (2009). *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide).* New Jersey: Technics Publications, LLC.

References

European Commission - ISA Programme. (2012). *ADMS Controlled Vocabularies*. Available at https://joinup.ec.europa.eu/svn/adms/ADMS_v1.00/ADMS_SKOS_v1.00.html

European Commission - ISA Programme. (2012). *SEMIC – 10 Rules for Persistent URIs*. Available at https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris

Publications Office of the EU. (2014). *Metadata Registry.* Available at http://publications.europa.eu/mdr/resource/authority/country/html/countries-eng.html#description.

ISO. (2014). *ISO/IEC 11179-1:2004 - Information technology - Metadata registries (MDR) - Part 1: Framework*. Available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35343

European Commission

joinup

Project Officers Vassilios.Peristeras@ec.europa.eu
Suzanne.Wigard@ec.europa.eu
Athanasios.Karalopoulos@ec.europa.eu

Visit our initiatives

ADMS ASSET DESCRIPTION METADATA SCHEMA

DCAT APPLICATION PROFILE FOR EUROPEAN DATA PORTALS

ADMS.SW

REGISTERED ORGANIZATION VOCABULARY

eGOVERNMENT CORE VOCABULARIES

CORE PUBLIC SERVICE VOCABULARY

CORE LOCATION VOCABULARY

CORE PERSON VOCABULARY

Get involved

Follow @SEMICeu on Twitter

Join the SEMIC group on LinkedIn

Join the SEMIC community on Joinup

ISA