CONSTELLATION
metadata registry

STIDS Conference
October 24, 2012

Gramm Richardson
U.S. Department of Defense
gpricha@tycho.ncsc.mil

Elli Schwarz
SRA International
eliezer_schwarz@sra.com

Country Code Sets

U.S. Government	
Census Schedule C	Used by the US Census Bureau as well as the Army Corps of Engineers.
Treasury International Capital Reporting	Designations identifying countries in data files on international portfolio capital movements reported to the US Treasury Department via the Treasury International Capital reporting system.
GSA Geographic Locator Codes	Used by US federal agencies for reporting data to the Federal Real Property Profile.
NGA Geopolitical Codes (and dependencies)	Codes for political entities in the NGA GEOnet Names Server (Formerly FIPS 10-4).

Country Code Sets

International Orgaizations	
ISO 3166-1, ISO 3166-2	Entities which are members of the UN or one of its specialized agencies and parties to the Statute of the International Court of Justice, or registered by the UN Statistics Division. Part 2 of the standard includes dependencies of the entities in Part 1.
International Olympic Committee	Codes identifying the National Olympic Committees/National Teams participating in the Olympics.
IANA Top-Level Domains	Top level domains used in web addresses like .com, .cc, .uk, .РФ, or .中国
UN FAO Geopolitical Ontology	AGROVOC, FAOSTAT, FAOTERM - code sets used for agricultural statistics and projects purposes.
UN M.49 Area Codes	Used by the United Nations for statistical purposes.

Country Code Sets

Industry	
ITU-T e.164	Recommendation that defines structure for telephone numbers, including country dialing codes.
ITU-T e.212	Defines the code used in the Mobile Country Code Portion of an IMSI (International Mobile Subscriber Identifier).
International Union of Railways	Standard numerical country coding for use in railway traffic. Used as the owner's code (3rd and 4th position) of a 12-digit wagon identification number.
International Civil Aviation Organization	Aircraft nationality marks based on the Chicago Convention on International Civil Aviation, as reported to ICAO by national administrations. Used as the prefix of an aircraft tail number.

Country Code Registry

Group by: Close Match UUID | Collapse All | Page 1 of 1 | 50 | Showing 1 - 10 of 10 results

Source ▾	Country Name	Code
⊟ **Close Match UUID:** _u84d2531c-3614-4c44-a98e-50ae49a0114f_ 10 Item(s)		
FIPS 10-4-alphabetic2	RUSSIA	RS
IANA-alphabetic	Russian Federation	.РФ
IANA-alphabetic	Russian Federation	.RU
ICAO Nationality Marks-alphanumeric	Russian Federation	RA
ISO 3166-1-alphabetic2	Russian Federation (the)	RU
ISO 3166-1-alphabetic3	Russian Federation (the)	RUS
ISO 3166-1-numeric3	Russian Federation (the)	643
ITU-T e.164-numeric	Russian Federation	7
ITU-T e.212-alphanumeric	Russian Federation	250
NGA Geopolitical Codes-alphabetic2	RUSSIA	RS

Page 1 of 1 | 50 | Showing 1 - 10 of 10 results

Managing Country Codes in a Metadata Registry

- Ambiguity in term "Country"
  - China/Hong Kong
  - Kosovo
  - Sovereign Military Order of Malta
  - European Union
  - Unknown/Reserved

Country Metadata Complexity

Existing approaches don't take some or all of these factors into account:

- There are many country code standards in use, each one with its own complement of values.
- Countries may have multiple names, in multiple languages, with multiple codes or code formats.
- Multiple countries may share the same code.
- Country lists may change over time
- Countries may have relationships with other countries (administrative, membership, etc)
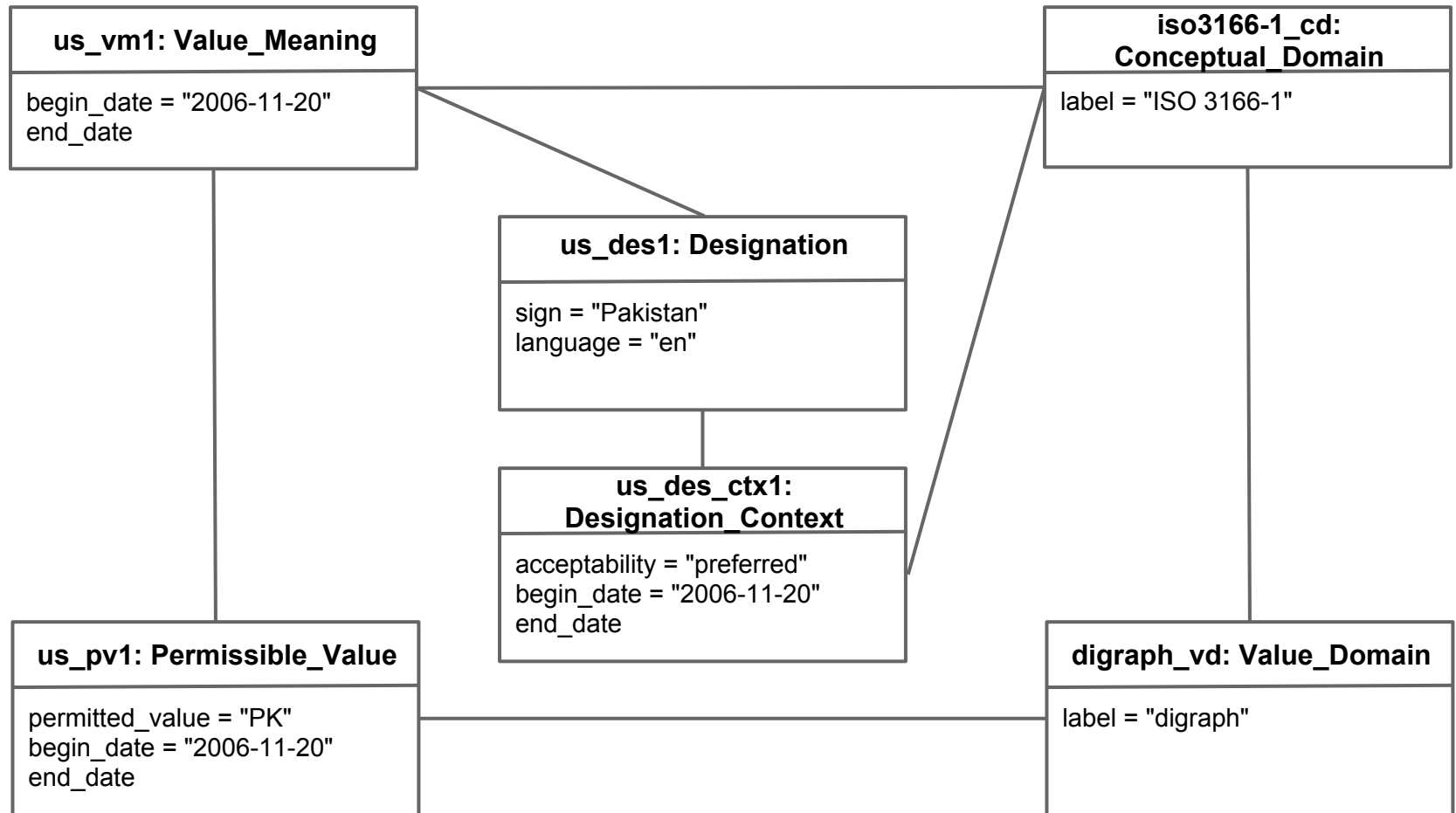
Country Metadata Complexity

- There are many country code standards in use, each one with its own complement of values.
  - *Register each country code set separately (none of the code sets are required to be privileged or canonical)*
  - *Map country entities across code sets*

- Countries may have multiple names, in multiple languages, with multiple codes or code formats.
  - *Have distinct classes for country, name, and code, each with its own properties.*
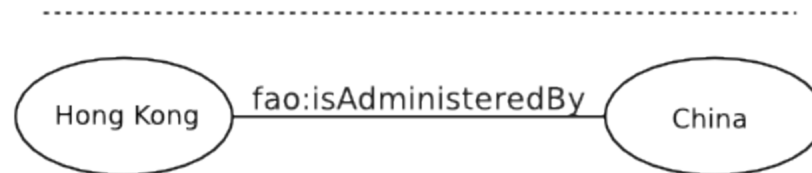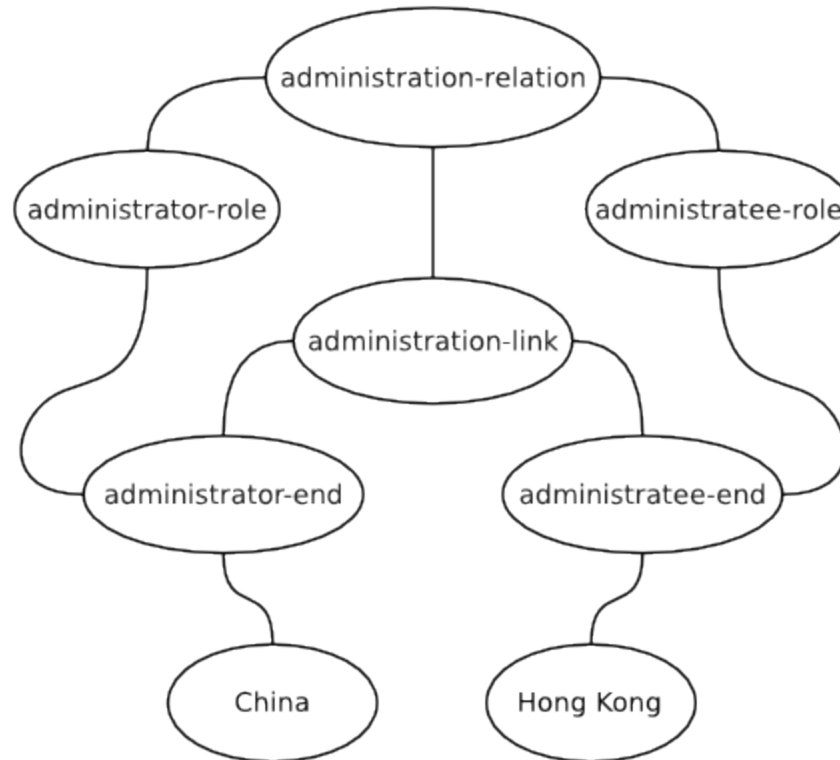
Country Metadata Complexity

- Country lists may change over time
  - *Keep track of all revisions separately for each of the above classes*
  - *Old data isn't removed*

- Countries may have relationships with other countries (administrative, membership, etc)
  - *Allow for arbitrary relationships among countries*

ISO/IEC 11179 in Constellation

us_vm1: Value_Meaning

begin_date = "2006-11-20"
end_date

**iso3166-1_cd:
Conceptual_Domain**

label = "ISO 3166-1"

us_des1: Designation

sign = "Pakistan"
language = "en"

**us_des_ctx1:
Designation_Context**

acceptability = "preferred"
begin_date = "2006-11-20"
end_date

us_pv1: Permissible_Value

permitted_value = "PK"
begin_date = "2006-11-20"
end_date

digraph_vd: Value_Domain

label = "digraph"

UML Object diagram showing an example of some instances of Constellation's ontology for 11179.

Relationships

Mapping Country Code Sets

- Statistical name similarity is used as a first pass to create lists of probable matching countries in two code sets
- Very high precision and recall, but we want perfect precision and recall
- Manually curated truth data is used to generate the mappings in our semi-automated process
- Manual matching is essential when names of a country differ significantly - Burma ⇄ Myanmar
- Sometimes, even manual evaluation can't give us perfect precision and recall
- Using a 11179 representation of SKOS relationships to associate entities
  - closeMatch
  - broadMatch/narrowMatch
  - exactMatch

Updating Code Set Metadata

- Need to account for
  - Create, Update, Deprecate
  - Country, Country Name, Country Code
- Changes are additive

Query results for CountryQueryByCodeDate.				
Group by: no group		Page 1 of 1 ▷▷ ▷| 7		Showing 1 - 7 of 7 results
codeSet ◆	**countryName** ◆	**code** ◆	**CodeBeginDate** ◆	**CodeEndDate** ◆
ISO 3166-1-alphabetic2	Sudan (the)	SD	2006-11-20Z	
ISO 3166-1-alphabetic3	Sudan (the)	SDN	2006-11-20Z	
ISO 3166-1-numeric3	Sudan (the)	736	2006-11-20Z	2011-08-09Z
ISO 3166-1-numeric3	Sudan (the)	729	2011-08-09Z	
ISO 3166-1-alphabetic2	South Sudan	SS	2011-08-09Z	
ISO 3166-1-alphabetic3	South Sudan	SSD	2011-08-09Z	
ISO 3166-1-numeric3	South Sudan	728	2011-08-09Z	
⌕ ↻		Page 1 of 1 ▷▷ ▷| 7		Showing 1 - 7 of 7 results

Updates to Libya's Name

Query results for DesignationQuery.

Group by: Source | Collapse All | Page 1 of 1 | 50 | Showing 1 - 8 of 8 results

Language	Name	Acceptability	Begin Date	End Date
⊟ **Source: ISO 3166-1** 8 Item(s)				
ar	Al Jamāhīrīyah al ʿArabīyah al Lībīyah	preferred	2006-11-20Z	2011-11-08Z
ar	Al Jamāhīrīyah al ʿArabīyah al Lībīyah	deprecated	2011-11-08Z	
ar	Lībiyā	preferred	2011-11-08Z	
en	Libyan Arab Jamahiriya (the)	preferred	2006-11-20Z	2011-11-08Z
en	Libyan Arab Jamahiriya (the)	deprecated	2011-11-08Z	
en	Libya	preferred	2011-11-08Z	
en	the Socialist People's Libyan Arab Jamahiriya	admitted	2006-11-20Z	2011-11-08Z
en	the Socialist People's Libyan Arab Jamahiriya	deprecated	2011-11-08Z	

Page 1 of 1 | 50 | Showing 1 - 8 of 8 results

Querying

11179 OWL ontology and SPARQL provides:
- Multiple views of the data
- Same queries for all code sets
- Grouping by matched countries
- Querying for different mappings of data
- Querying by code, name, or dates
- Querying for names in multiple languages or acceptabilities
- Data analysis (Crosswalks, diffs, redundancy, etc.)
- SPARQL 1.1 is instrumental

Future Directions

- ## Other metadata artifacts
  - Thesauri
  - Rules
- ## Using reusable SPARQL functions (SPIN)
- ## Metamodel extensions
  - Additional attributes for metadata objects presented at the Metadata Standards Open Forum