Metadata practices to support discovery in Geospatial Platform/data.gov

Metadata collection has been ongoing for many years now in the geospatial field. With the publication of the initial Content Standard for Digital Geospatial Metadata (CSDGM) in 1994 and its Version 2 in 1998, and globally with the international metadata standard ISO 19115 in 2003, and its XML form, ISO TS19139, in 2007. The primary objective of these standards is to provide contextual documentation of geospatial resources – data, services, documents, and applications that enable fitness-for-use evaluations, and facilitate direct online access to the described resources.

The Geospatial One-Stop and the Geospatial Platform successor catalog, executed in partnership with data.gov, have as a priority the documentation of online accessible data and information products, specifically the location of data sets and applications for download and Web service URLs for real-time visualization and data access or processing. This document identifies best practices of metadata creation and publishing that will lead to more successful propagation of metadata and ease of resource discovery and access through cataloguing systems like data.gov and the Geospatial Platform.

URLs to metadata should be to online resources. Resources (data, services, applications) that are described in either CSDGM or ISO metadata must include URLs that take a user to the online resource. This is a baseline expectation of the data.gov environment and one we must follow in the shared catalog. Links should provide direct access to the resource wherever possible. Metadata records that only include references to websites or HTML pages where users must re-initiate search might not be harvested into the new catalog. As both metadata standards allow for multiple resource links to be placed in metadata, this should be encouraged to allow users to directly access multiple facets of the data: web service, download, documentation, browse graphic.

Use the appropriate metadata structure to describe the download or access URL. In CSDGM metadata, links to resources should be in the Distribution section of the standard using the following XML substructure, where Format Name (formname) should describe the specific format being made available in this group. The paired online URL is in the Network Resource (networkr) element. Multiple resources may exist in a metadata document for links to different formats. Online Linkage (onlink) is sometimes used to store actionable URLs, but is not recommended for encoding the distribution linkage.

```
<digform>
   <digtinfo>
      <formname>shapefile</formname>
   </digtinfo>
   <digtopt>
      <onlinopt>
         <computer>
            <networka>
```

```
            <networkr>http://someagency.gov/data/prod_234.zip
                        </networkr>
          </networka>
        </computer>
      </onlinopt>
    </digtopt>
</digform>
```

(I need the equivalent ISO structure for parity here)

Metadata will be indexed and properly classified as to resource type where format name is used. This clarifies the information resource payload to be accessed at the link. Data, service, and application resource types can be detected by the proper use of the values described at the end of this document for Format Name – paired with the URL it is describing.

Include terms found in published taxonomies in your metadata. In ISO Metadata, there is a structure allowing for the identification of one or more Topic Categories that best characterize your data or application domain. These are very high-level thematic bins, but are excellent for providing a quick triage of available data when indexed in the data.gov catalog. Use the ISO 19115/19139 structure for ISO Topic Category or the following example structure when using CSDGM metadata:

```
<keywords>
        <themekt>ISO 19115 Topic Category</themekt>
        <themekey>biota</themekey>
        <themekey>environment</themekey>
</keywords>
```

The valid values for ISO Topic Category keywords are:

biota
boundaries
climatologyMeteorologyAtmosphere
economy
elevation
environment
farming
geoscientificInformation
health
imageryBaseMapsEarthCover
inlandWaters
intelligenceMilitary
location
oceans
planningCadastre
society
structure

transportation
utilitiesCommunication

When adding these theme codes, specify "ISO 19115 Topic Category" as the theme keyword thesaurus.

For example, if your community uses the Global Change Master Directory Science Keywords be sure to reference the source in the Theme Keyword Thesaurus field (themekt = GCMD) in your metadata with a valid string (value) found in the following file:
http://gcmdservices.gsfc.nasa.gov/static/kms/sciencekeywords/sciencekeywords.csv

Publish only your original metadata/data. Many legacy collections of metadata exist where entries from partner agencies are included within the catalog. This may have been done to collate available resources into a community catalog. When harvested into a common index, this practice may leads to duplication of metadata. Worse, it may imply ownership or authorship of data from other sources.

Data.gov takes a focused view that resources being published should be 'original' to the publishing agency. When committing data to data.gov, a publisher is asserting that the data conform to the data quality guidelines of the publishing organization and that they are authors of such data. If an agency participated in the development of the data set (sponsorship, co-development, collection, modification, re-processing) then the metadata may be published by the agency to catalog.data.gov with clear attribution to the source and participating organizations as a co-authored or derivative product. Metadata for data not originating in an agency should not be registered with the data.gov catalog.

Titles should be descriptive. Following good library and publications practice, titles of metadata documents should be declarative so that when reviewed, the user has a good sense of the information context being promoted. Titles must also be unique – having a dozen results with the identical title is not very helpful to the end user and there are features within the results that should be elevated to the title to distinguish them. Good titles will include the topic, and where relevant, the range of dates and geography associated with the data product being documented – What? Where? When?. Where resources have local, regional, national, or global coverage, it should be apparent in the title. If the metadata describes a modification of data from another source, that also should be noted in the title. Never use a filename as the title of a metadata record! Although titles should not try to replace an Abstract or more lengthy description, they should not be terse.

Collections or series of homogeneous data need collection-level and member-level metadata records. Homogeneous series data are all of the same content, share most of the same metadata values, and might only vary in terms of content date and geographic extent. Examples include satellite imagery

repositories, or data product series individually available for download. This type of collection management is not applicable to most heterogeneous collections where every record should be indexed and is unique relative to its peers within the collection.

To support homogeneous collections, the new catalog.data.gov solution supports the ability to distinguish a 'parent' metadata record that contains the basic information common to the full series. The first resource description link (under the distribution section) should be a URL to the Web Accessible Folder or CSW GetRecords URL containing the series metadata records. These 'child' records are separately indexed and made discoverable to the user after finding the parent record in general search. The parent record should not be located with the children in the folder or its subfolders.

Dates of information content should be documented. Both CSDGM and ISO metadata include metadata fields to store dates related to the metadata and the information content itself. In CSDGM, the Time Period of Content structure carries the beginning and ending date of the data, expressed using ISO 8601 Dates in the form of YYYYMMDD or YYYY-MM-DD. The word 'present' may also be used in lieu of an end-date for continuously collected data.

In ISO metadata, the following structure should be used, by example, using the same date format:

```
<gml:TimePeriod gml:id="boundingTemporalExtent">
   <gml:description>ground condition</gml:description>
   <gml:beginPosition>1999-11-03</gml:beginPosition>
   <gml:endPosition>2012-12-30</gml:endPosition>
</gml:TimePeriod>
```

The indeterminatePosition attribute can be used in time positions. This attribute is often used to document unknown and present dates. The valid values for indeterminatePosition are "unknown", "after", "before", and "now". The use of "now" would parallel the use of "present" in CSDGM:

```
   <gml:endPosition indeterminatePosition="now"/>
```

Publication date can also be encoded in the ISO Identification Section as a useful date for search on information products, as per the following example:

```
<gmd:date>
  <gmd:CI_Date>
    <gmd:date>
      <gco:Date>2010-06-04</gco:Date>
    </gmd:date>
    <gmd:dateType>
      <gmd:CI_DateTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gm
xCodelists.xml#CI_DateTypeCode"
```

```
      codeListValue="publication"
      codeSpace="002">publication</gmd:CI_DateTypeCode>
    </gmd:dateType>
  </gmd:CI_Date>
</gmd:date>
```

Supplement: Recommended format and protocol codes to be used in the 'Format Name' field for proper type classification in data.gov

Known raster geospatial data formats (resource type: data) include the following:

Description	'Format Name' Value
Arc/Info ASCII Grid	AAIGrid
ADRG/ARC Digitilized Raster Graphics (.gen/.thf)	ADRG
Arc/Info Binary Grid (.adf)	AIG
Magellan BLX Topo (.blx, .xlb)	BLX
Bathymetry Attributed Grid (.bag)	BAG
Microsoft Windows Device Independent Bitmap (.bmp)	BMP
VTP Binary Terrain Format (.bt)	BT
USGS LULC Composite Theme Grid	CTG
Spot DIMAP (metadata.dim)	DIMAP
DODS / OPeNDAP	OPeNDAP
USGS DOQ	DOQ
Military Elevation Data (.dt0, .dt1, .dt2)	DTED
Arc/Info Export E00 GRID	E00GRID
ERDAS Compressed Wavelets (.ecw)	ECW
Erdas Imagine Raw	EIR
ENVI .hdr Labelled Raster	ENVI
Epsilon - Wavelet compressed images	EPSILON
ERMapper (.ers)	ERS
Envisat Image Product (.n1)	ESAT
EOSAT FAST Format	FAST
Graphics Interchange Format (.gif)	GIF
WMO GRIB1/GRIB2 (.grb)	GRIB
GRASS Rasters	GRASS
GRASS ASCII Grid	GRASSASCIIGrid
Golden Software Surfer 7 Binary Grid	GS7BG
TIFF / BigTIFF / GeoTIFF (.tif)	GeoTiff
GXF - Grid eXchange File	GXF
Hierarchical Data Format Release 4 (HDF4)	HDF4
Hierarchical Data Format Release 5 (HDF5)	HDF5
Erdas Imagine (.img)	HFA
Image Display and Analysis (WinDisp)	IDA
ILWIS Raster Map (.mpr,.mpl)	ILWIS
Intergraph Raster	INGR
IRIS	IRIS
USGS Astrogeology ISIS cube (Version 2)	ISIS2

USGS Astrogeology ISIS cube (Version 3)	ISIS3
JAXA PALSAR Product Reader (Level 1.1/1.5)	JAXAPALSAR
Japanese DEM (.mem)	JDEM
JPEG JFIF (.jpg)	JPEG
JPEG-LS	JPEGLS
JPEG2000 (.jp2, .j2k)	JPEG2000
JPIP (based on Kakadu)	JPIPKAK
KMLSUPEROVERLAY	KMLSUPEROVERLAY
NOAA Polar Orbiter Level 1b Data Set (AVHRR)	L1B
Erdas 7.x .LAN and .GIS	LAN
FARSITE v.4 LCP Format	LCP
Daylon Leveller Heightfield	Leveller
MBTiles	MBTiles
OziExplorer .MAP	OZIMAP
Vexcel MFF	MFF
Vexcel MFF2 (HKV)	MFF2
MG4 Encoded Lidar	MG4Lidar
Multi-resolution Seamless Image Database	MrSID
Meteosat Second Generation	MSG
EUMETSAT Archive native (.nat)	MSGN
NLAPS Data Format	NDF
NOAA NGS Geoid Height Grids	NGSGEOID
NITF (.ntf, .nsf, .gn?, .hr?, .ja?, .jg?, .jn?, .lf?, .on?, .tl?, .tp?, etc.)	NITF
NetCDF	netCDF
PCI .aux Labelled	PAux
PCI Geomatics Database File	PCIDSK
Geospatial PDF	PDF
NASA Planetary Data System	PDS
Portable Network Graphics (.png)	PNG
PostGIS Raster (previously WKTRaster)	PostGISRaster
Rasterlite - Rasters in SQLite DB	Rasterlite
Raster Matrix Format (*.rsw, .mtw)	RMF
Raster Product Format/RPF (CADRG, CIB)	RPFTOC
RadarSat2 XML (product.xml)	RS2
Idrisi Raster	RST
SAGA GIS Binary format	SAGA
SAR CEOS	SAR_CEOS
ArcSDE Raster	SDERASTER
USGS SDTS DEM (*CATD.DDF)	SDTS

SGI Image Format	SGI
Snow Data Assimilation System	SNODAS
Standard Raster Product (ASRP/USRP)	SRP
SRTM HGT Format	SRTMHGT
Terragen Heightfield (.ter)	TERRAGEN
EarthWatch/DigitalGlobe .TIL	TIL
TerraSAR-X Product	TSX
USGS ASCII DEM / CDED (.dem)	USGSDEM
ASCII Gridded XYZ	XYZ
ZMap Plus Grid	ZMap

Known GIS Vector data formats (resource type: data) include the following:

Description	'Format Name' Value
Aeronav FAA files	AeronavFAA
Arc/Info .E00 (ASCII) Coverage	E00
Arc/Info Generate	ARCGEN
AutoCAD DWG	DWG
AutoCAD DXF	DXF
CouchDB / GeoCouch	CouchDB
ElasticSearch	ElasticSearch
ESRI FileGDB	FileGDB
ESRI Personal GeoDatabase	PGDB
ESRI Shapefile	Shapefile
GeoJSON	GeoJSON
Geomedia .mdb	Geomedia
GeoRSS	GeoRSS
GML	GML
GMT	GMT
GPS Exchange Format GPX	GPX
OGC KML (.kml, .kmz)	KML
Mapinfo File (.mif, .tab)	MapInfo
Microstation DGN	DGN
Vector Product Format	VPF
OpenStreetMap XML and PBF	OSM
PCI Geomatics Database File	PCIDSK
Geospatial PDF	PDF
PostgreSQL SQL dump	PGDump
S-57 (ENC)	S57
SDTS	SDTS
SEG-P1 / UKOOA P1/90	SEGUKOOA

SEG-Y	SEGY
SQLite/SpatiaLite	SQLite
SUA	SUA
SVG	SVG
U.S. Census TIGER/Line (non-shapefile)	TIGER
X-Plane/Flightgear aeronautical data	XPLANE

Known tabular data formats (resource type: data) that may contain spatial columns:

Description	'Format Name' Value
Comma Separated Value (.csv)	CSV
Google Fusion Tables	GFT
Open Document Spreadsheet	ODS
MS Excel format	XLS
MS Office Open XML spreadsheet	XLSX

Known service protocols (resource type: Service) that can be encoded in Format Name if describing a service and not a data set-specific endpoint:

Description	'Format Name' Value
OGC Web Feature Service	WFS
OGC Web Coverage Service	WCS
OGC Web Map Service	WMS
OGC Web Processing Service	WPS
OGC Catalog Service for the Web	CSW
Open-source Project for a Network Data Access Protocol (OPeNDAP, THREDDS)	OPeNDAP
ArcGIS REST Service API	ArcREST

Identified formats for applications include the following:

Mobile application – iOS (iPad, iPhone, iTouch)	iOS
Mobile application – Android	APK
Mobile application – jQuery, HTML5	jQuery
Web application - general	HTML5
Web application – Flash	Flash
Web application – Flex	Flex
Windows application	EXE