

EPA Guidance for Creating Homogeneous Collections

Contents

EPA Guidance for Creating Data.gov Homogeneous Collections	1
Background	1
Methods to Group Metadata Records.....	1
Data.gov Collections	1
Instructions for Producing a Collection	2
FGDC CSDGM Metadata	3
Parent Record	3
Child Records	4
ISO 19115 Metadata	4
Parent Record	4
Child Records	5
Non-Geo Metadata.....	7
Non-geospatial Metadata.....	7
Viewing Homogenous Collections in the EDG	8
Viewing Homogenous Collections at Data.gov.....	9

Background

Methods to Group Metadata Records

In the EPA’s Environmental Dataset Gateway (EDG) there are two ways to group metadata records – collections and compilations.

Compilations refer to a flexible grouping of metadata records. This grouping can be according to any practical line of reasoning – common subject matter, business case, application, project, theme, etc. Historically, this type of grouping was called a collection. However, recently GSA and OMB defined a collection much more narrowly, the historic concept of a flexible EDG collection has been renamed a *compilation* to draw a distinction between the two. Compilations only exist within the EDG, they are not represented in Data.gov. Guidance documents for creating and managing EDG compilations may be found at <https://edg.epa.gov/>.

The focus of this guidance document is on how to create EDG and Data.gov collections.

Data.gov Collections

Project Open Data (<https://project-open-data.cio.gov>) defines a collection as:

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.

A collection of this type is comprised of a single parent metadata record that describes the collection as a whole, and child records which contain embedded references to the parent. When the parent/child relationship is embedded in the metadata itself, rather than stored as a linkage in a metadata catalog, it allows that relationship to persist as the metadata are harvested and aggregated from catalog to catalog. Examples of homogenous collections of EPA data include Toxic Release Inventory (TRI) data released by year and state, or Re-Powering Alternative Energy data released by EPA Region.

Comparison of Compilations versus Collections:

Compilation	Collection
Only in EDG	In EDG and Data.gov
Made up of heterogeneous or homogeneous records	Made up of homogenous records
Example: All data used in the EJSCREEN application	Example: TRI data released by year and state

Instructions for Producing a Collection

The EDG supports two core geospatial metadata formats – Geospatial (FGDC CSDGM¹ and ISO 19115²) and one non-geospatial format (Project Open Data, or POD). Because each metadata format is handled differently at Data.gov, we strongly recommend that all metadata records in a collection, including the parent record, are in the same format. The process for creating homogenous collections in each of these three formats is outlined in the sections below.

The first step for creating collections in all metadata formats is to identify or create a parent metadata record that represents the entire homogeneous collection; this parent record must exist and have a valid Universal Unique Identifier (UUID) before proceeding. When metadata records are first contributed to the EDG they are assigned a UUID³. To find the UUID of a metadata record in the EDG, simply perform a search for that record. Once the record is located, open the record's Details page. The URL in the browser address bar contains the UUID as the final parameter (uuid=%7B980A5659-9D0F-4A60-9183-3BBB49CD5CD6%7D), see Figure 1.

¹ [Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata](#)

² [International Organization for Standardization \(ISO\) standard number 19115 - Geographic information – Metadata](#)

³ If your metadata records are not in the EDG and you wish to generate your own UUID ahead of time, you may use any UUID generator (<https://www.uuidgenerator.net>, for example) and follow the instructions in the [EDG Metadata recommendations](#) to embed the UUID prior to harvesting.



Figure 1 - Location of the UUID in the EDG Metadata Details URL

In the example URL below, the UUID is B6FE56EE-3D28-4B5C-ABF0-D3B0B9E9DF87. Please note that while the EDG consistently wraps curly braces {} around the UUID, that when embedding UUIDs in metadata records the curly braces should be omitted. It's also not uncommon for the curly braces to be replaced by their URL encodings: %7B and %7D. Either way, the braces don't belong in the metadata.

<https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid={B6FE56EE-3D28-4B5C-ABF0-D3B0B9E9DF87}>

FGDC CSDGM Metadata

Parent Record

The first step in creating a collection using FGDC CSDGM metadata is to ensure that the UUID is appropriately embedded in the parent record – in two separate places. Neither the EPA Metadata Editor (EME 3.2.1) nor the Esri ArcCatalog metadata editor facilitate editing of these elements, so it is necessary to edit the child metadata records by hand in a text editor (or some other XML editing tool).

The first location is an Esri custom tag “<Esri>” (metadata/Esri/PublishedDocID) that ensures UUIDs are preserved across harvests. The second location (metadata/idinfo/citation/citeinfo/othcit) follows guidance provided by the FGDC regarding “Documentation of Collection in Geospatial Metadata In Support of Project Open Data”⁴. In the example below, the elements highlighted in red are the two locations where the UUID of the parent record needs to be embedded in itself:

```
<metadata>
  <Esri>
    <PublishedDocID>B6FE56EE-3D28-4B5C-ABF0-D3B0B9E9DF87</PublishedDocID>
  </Esri>
  <idinfo>
    <citation>
      <citeinfo>
        <origin>US Environmental Protection Agency</origin>
        <title>EJSCREEN Data--2015 Public Release</title>
        <pubinfo>
          <publish>U.S. Environmental Protection Agency, Headquarters</publish>
          <pubplace>Washington, DC</pubplace>
        </pubinfo>
        <pubdate>20150716</pubdate>
      </citeinfo>
    </citation>
  </idinfo>
</metadata>
```

⁴ Still under review, will provide an online linkage when available.

```

        <othcit>B6FE56EE-3D28-4B5C-ABF0-D3B0B9E9DF87</othcit>
        <onlink>ftp://newftp.epa.gov/EJSCREEN/</onlink>
    </citeinfo>
</citation>

```

Child Records

Once the parent record is complete, a child metadata record may reference a parent through the use of the Larger Work Citation (*lworkcit*). Since a *lworkcit* is a simple container for a *citeinfo*, the most straightforward method is to copy the entire *citeinfo* section from the parent record above and paste it into the child record within the *lworkcit* element. Below is an example of how the *lworkcit* section fits into the main citation section of a CSDGM record, with the critical section highlighted in red:

```

<metadata>
  <idinfo>
    <citation>
      <citeinfo>
        <origin>US Environmental Protection Agency</origin>
        <title>EJSCREEN Demographic Indicators 2015 Public</title>
        <pubinfo>
          <publish>U.S. Environmental Protection Agency, Headquarters</publish>
          <pubplace>Washington, DC</pubplace>
        </pubinfo>
        <pubdate>20150716</pubdate>
        <lworkcit>
          <citeinfo>
            <origin>US Environmental Protection Agency</origin>
            <title>EJSCREEN Data--2015 Public Release</title>
            <pubinfo>
              <publish>U.S. Environmental Protection Agency, Headquarters</publish>
              <pubplace>Washington, DC</pubplace>
            </pubinfo>
            <pubdate>20150716</pubdate>
            <othcit>B6FE56EE-3D28-4B5C-ABF0-D3B0B9E9DF87</othcit>
            <onlink>ftp://newftp.epa.gov/EJSCREEN/</onlink>
          </citeinfo>
        </lworkcit>
      </citeinfo>
    </citation>
    <onlink>http://ejscreen.epa.gov/arcgis/rest/services/ejscreen/Demographic_Indicators_2015_Public/MapServer</onlink>
    <onlink>ftp://newftp.epa.gov/EJSCREEN/</onlink>
  </citeinfo>
</citation>

```

Any *citeinfo* section must contain *origin*, *pubdate*, and *title* elements per the CSDGM specification, which is why those elements are included within the *lworkcit* section, despite their apparent redundancy. The *othcit* element is critical to the creation of the collection. Other citation elements, such as *pubinfo* and *onlink*, may be omitted from the *lworkcit* section if they are seen as overly redundant. The finished records may then be contributed to the EDG via the standard procedure. Please see the last section in this document, “Viewing Homogenous Collections in the EDG” to verify successful linkages.

ISO 19115 Metadata

Parent Record

ISO contains a simple element at the root for embedding the unique identifier of any metadata record, unlike FGDC CSDGM. This element appears like this:

```
<gmd:MD_Metadata>
  <gmd:fileIdentifier>
    <gco:CharacterString>DABABBD-0A2F-48F9-BC78-DD748B588FC5</gco:CharacterString>
  </gmd:fileIdentifier>
```

and occurs at the top of the document, right after the MD_Metadata or MI_Metadata tag. More information about this element is available here:

https://geo-ide.noaa.gov/wiki/index.php?title=MI_Metadata

In the standard ArcCatalog metadata editor this element appears in the Metadata Details pane, and ArcCatalog contains a tool to generate a new identifier if one does not already exist – the “Create” button on the right in Figure 2.

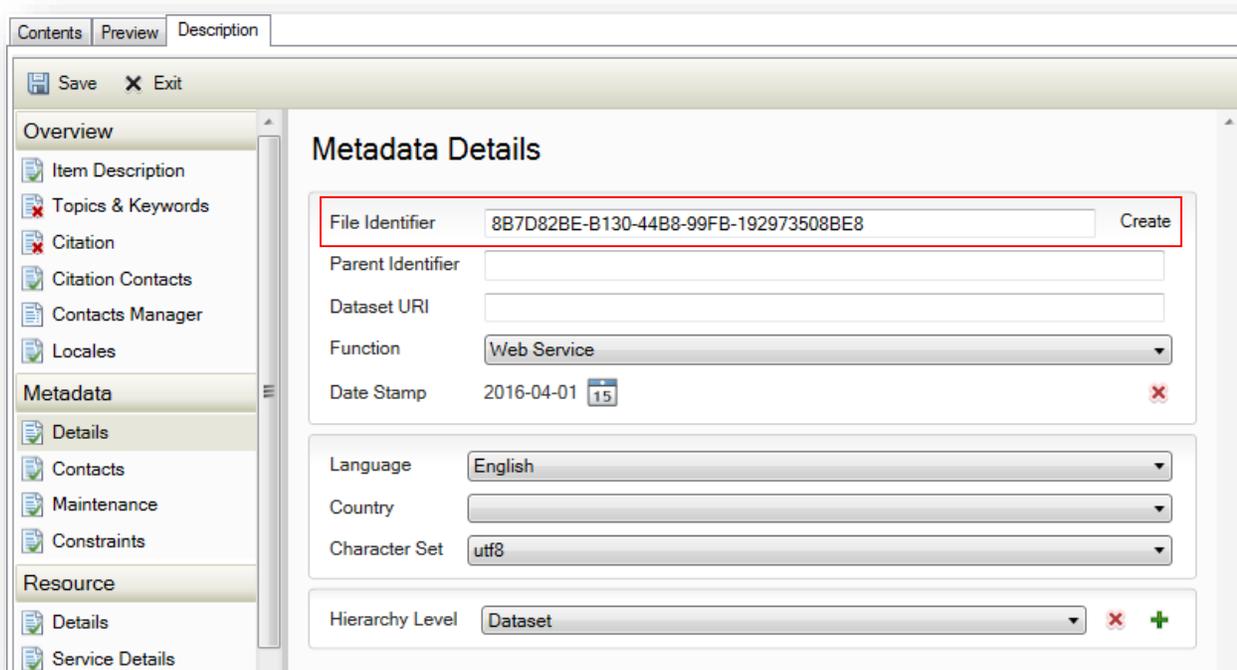


Figure 2 - File Identifier location in ArcCatalog Metadata Editor

This is appropriate in the creation of a new parent record, otherwise, the existing UUID of the parent record should be cut and pasted from the EDG URL as described above.

Child Records

It may be tempting to assume that the Parent Identifier element (seen just below File Identifier in the ArcCatalog interface in Figure 2) is the correct way for a child to reference a parent in a collection, but according to guidance from the FGDC, this Parent Identifier is used to describe a parent/child relationship between metadata documents that together comprise a whole metadata document (it's not uncommon for ISO metadata sections, such as Lineage or Entity/Attribute, to be split out into separate but linked standalone documents). Since the goal of this exercise instead is to describe a collection of complete metadata records the appropriate method is to use the *aggregationInfo* section with *largerWorkCitation* as the *associationTypeCode*:

```
<gmd:MD_Metadata>
```

...

```

<gmd:identificationInfo>
...
</gmd:resourceConstraints>
<gmd:aggregationInfo>
  <gmd:MD_AggregateInformation>
    <gmd:aggregateDataSetIdentifier>
      <gmd:MD_Identifier>
        <gmd:code>
          <gco:CharacterString>DABABBBB-0A2F-48F9-BC78-DD748B588FC5
          </gco:CharacterString>
        </gmd:code>
      </gmd:MD_Identifier>
    </gmd:aggregateDataSetIdentifier>
    </gmd:associationType>
    <gmd:DS_AssociationTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#DS_AssociationTypeCode"
codeListValue="largerWorkCitation"
codeSpace="ISOTC211/19115">largerWorkCitation</DS_AssociationTypeCode>
    </gmd:associationType>
  </gmd:MD_AggregateInformation>
</gmd:aggregationInfo>
<gmd:spatialRepresentationType>

```

In ArcCatalog this is located in the "References" tab (Figure 3):

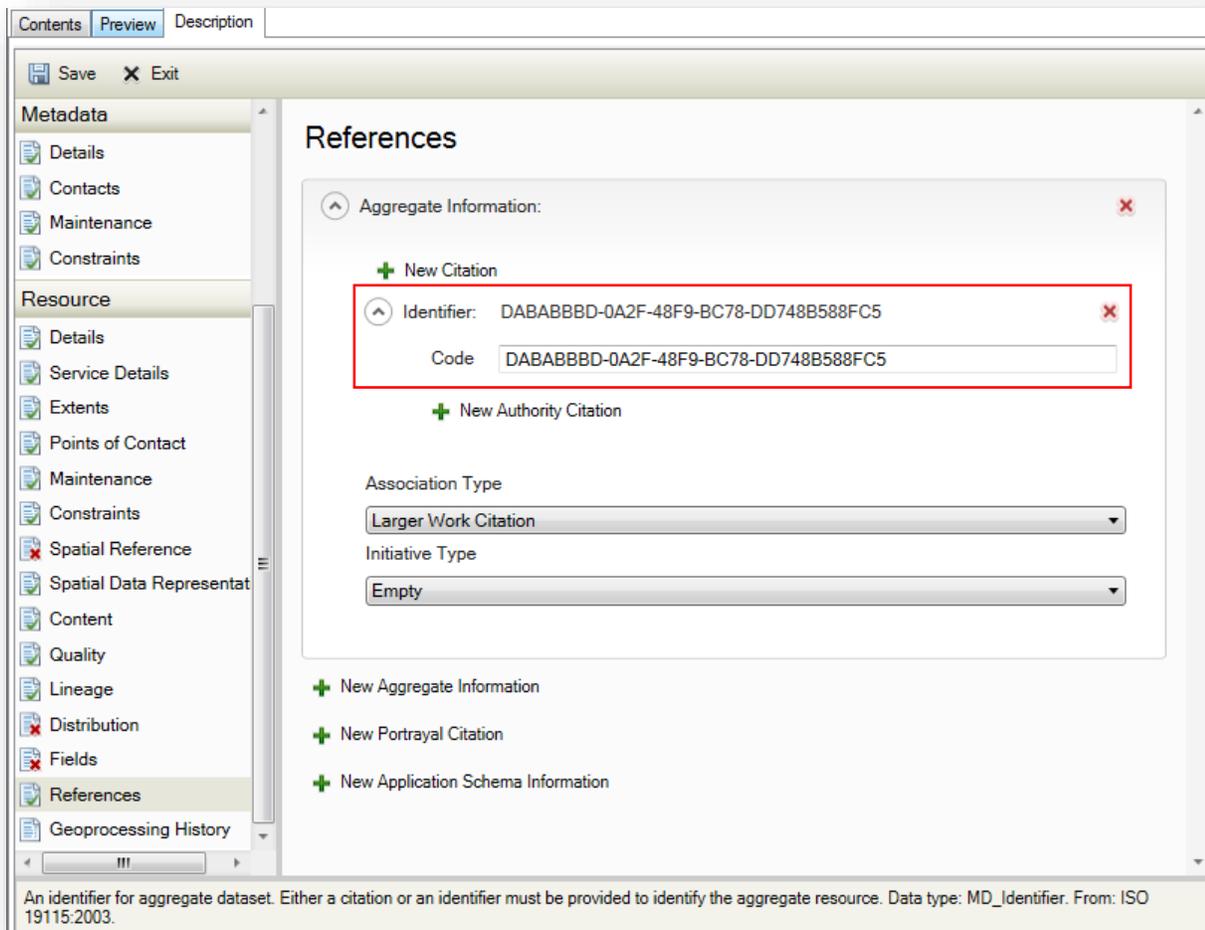


Figure 3 - Collection Parent Identifier in ISO-19115 Metadata

Non-Geo Metadata

Non-geospatial Metadata

Non-geospatial metadata records follow the Project Open Data (POD) schema, outlined here:

<https://project-open-data.cio.gov/v1.1/schema/>

EDG stewards comfortable authoring data.json files may follow this guidance and use the *identifier* element to specify a UUID for a parent record, and the *isPartOf* element to reference that parent identifier from child records.

Most EDG stewards prefer managing their non-geospatial metadata in an Excel spreadsheet that can be converted into and out of POD json. If you wish to create or edit your records in such a spreadsheet, please contact the EDG team at edg@epa.gov, and we will send you the latest version of your metadata. In this spreadsheet, each row represents a distinct metadata record. You may find the unique identifier of the parent record in the Unique Identifier column (or if you are authoring a new parent record, populate this column with a newly generated UUID using a generator website such as <https://www.uuidgenerator.net>). Copy and paste this parent UUID into the "Collection" column for all of the child records. We anticipate additional tools being available for editing non-geo records in the near future.

Viewing Homogenous Collections in the EDG

Homogenous collections may be viewed in the EDG by visiting the “Details” page of a record that is a member of a collection, and clicking on the “Relationships” link near the top. If a metadata record participates in a collection as a child, then selecting “Child of” will display the parent record (Figure 4). Similarly, if a record participates in a collection as a parent, then selecting “Parent to” will display all of the children (Figure 5).

The screenshot shows the EPA Environmental Dataset Gateway (EDG) interface. At the top, the EPA logo and the title "Environmental Dataset Gateway (EDG)" are displayed, with the tagline "Connecting EPA's Environmental Resources". Below this is a navigation menu with links for HOME, ABOUT, SEARCH, BROWSE, DATA, REUSE, and RESOURCES. The current page is titled "Details" and "Review Relationships". The main heading is "EPA FRS Facilities Single File CSV Download for the State of Nebraska". Below the heading, a sub-heading states: "This page lets you browse metadata by relationships with currently viewed metadata." On the left side, there is a tree view under "Resource" with options: "Child of", "Parent to", and "Others from this publisher". The "Child of" option is selected. The main content area shows "1 results" and "Showing 1-1". The result is titled "EPA Geospatial Data Download: Facility and Site Information" and contains a description: "Contains information about facilities or sites subject to environmental regulation, including key facility information along with associated environmental interests for use in mapping and reporting applications." Below the description are links for "Open Website Details Metadata" and "Internet REST Links: GEORSS HTML FRAGMENT KML" and "Intranet REST Links: GEORSS HTML FRAGMENT KML".

Figure 4 - Example of a child record showing its relationship with its parent in the EDG

The screenshot shows the EPA Environmental Dataset Gateway (EDG) interface. At the top, the EPA logo and the title "Environmental Dataset Gateway (EDG)" are displayed, with the tagline "Connecting EPA's Environmental Resources". Below this is a navigation menu with links for HOME, ABOUT, SEARCH, BROWSE, DATA, REUSE, and RESOURCES. The current page is titled "Details" and "Review Relationships". The main heading is "EPA Geospatial Data Download: Facility and Site Information". Below the heading, a sub-heading states: "This page lets you browse metadata by relationships with currently viewed metadata." On the left side, there is a tree view under "Resource" with options: "Child of", "Parent to", and "Others from this publisher". The "Parent to" option is selected. The main content area shows "119 results" and "Showing 1-10" with pagination links "1 2 3 4 5 > Last". The results are listed as follows: "Facility Registration System (FRS) Widget" with description "The Facility Registry System (FRS) widget returns facilities in a user-specified area of interest that report to more than one EPA program system as reported by the Facility Registry System" and links "Website Details Metadata"; "EPA FRS Facilities Single File CSV Download for the State of North Dakota" with description "The Facility Registry System (FRS) identifies facilities, sites, or places subject to environmental regulation or of environmental interest to EPA programs or delegated states. Using vigorous verification and data management procedures, FRS integrates fac..." and links "Website Details Metadata"; and "EPA FRS Facilities Single File CSV Download for the State of Nebraska".

Figure 5 - Example of a parent record showing its relationship with its children in the EDG

Viewing Homogenous Collections at Data.gov

The key motivation for grouping homogenous metadata records into collections in data.gov is to reduce potential clutter in search results. To this end, metadata records that participate in a collection as children will not appear in search results at data.gov – only the parent record will appear, and an icon will be displayed next to the title indicating that it represents a collection (Figure 6 - Example of a Collection in the Data.gov Search Results (Figure 6)).

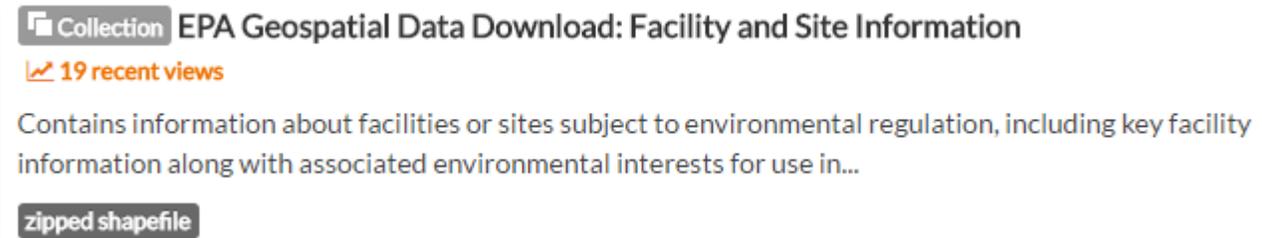


Figure 6 - Example of a Collection in the Data.gov Search Results

On the page that shows the full parent metadata record, a prominent link is available to “Search datasets within this collection” (Figure 7). Clicking this link will show all the child records and allow a user to perform additional searches for specific records within the collection.



EPA Geospatial Data Download: Facility and Site Information

Metadata Updated: Jan 23, 2016

Contains information about facilities or sites subject to environmental regulation, including key facility information along with associated environmental interests for use in mapping and reporting applications.

Access & Use Information

- Public:** This dataset is intended for public access and use.
- License:** See this page for license information.

Publisher

U.S. EPA Office of Environmental Information (OEI) - Office of Information Collection (OIC)

Contact

David G. Smith

Share on Social Sites

Google+

Twitter

Facebook

Collection

This dataset is a collection of other datasets.

[Search datasets within this collection](#)

Downloads & Resources

 **Data describes facilities currently regulated ...**  48 views
EPAShapefileDownload.zip

[Download](#)

Figure 7 - Link from parent to children on data.gov metadata page