EPA Metadata Crosswalk Implementation

Torrin Hultgren, Innovate!, Inc
hultgren.torrin@epa.gov
A compressed history

• Baseline Architecture
• The link between ISO and Project Open Data
• EPA Metadata Editors
• How we produce data.json
• Schema 1.0 compliance and pivot to Schema 1.1
• Path forward
Baseline Architecture

- Federated metadata management by a network of stewards
- Metadata produced using the EPA Metadata Editor
- Harvested into central metadata catalog (Esri GeoPortal Server) and from there to data.gov
- Geospatial community uses Esri stack and CSDGM
- Non-geo community used DMS and spreadsheets
DMS + POD = ISO?

• Previously implemented custom support for now-obsolete data.gov DMS format – strong motivation not to make the same mistake twice
• Initially GeoPortal could produce dcat.json, but not consume it
• CSDGM not a great fit for Project Open Data (POD) schema, ISO on our roadmap
• Conclusion – map POD to ISO, accelerate ISO adoption
• EME 3.2 extremely successful metadata editor with both ArcCatalog integration and standalone functionality, but hardwired to Microsoft Access database and CSDGM schema
• ISO Support required near complete rewrite
• Development launched with goal of supporting fields necessary for data.gov compliance, full ISO editing implementation to follow
Spreadsheet Editor

- Authored Excel spreadsheet with POD guidance, valid values as dropdowns and example records
- On receipt of spreadsheet, save as CSV, run Python script to map values to ISO xml template – harvest as ISO
- Great for batch editing, clunky in most other ways
- Map to ISO functional, but authored in isolation, and many fields required creative interpretation of ISO
• Summer 2014: beta tests of EME 4.0 favorable, but geospatial community refused to convert records to ISO until Data Quality and Feature Catalog sections could be part of output (ISO 19115-3, Spring 2015)
• Meant November 30th 2014 deadline of full POD schema 1.0 compliance had to be met with CSDGM records
• Produced EME 3.2.1 for minimal POD compliance with CSDGM records
Esri GeoPortal and POD

• Permits flexible mapping of elements from standard metadata formats (CSDGM, ISO, Dublin Core) to DCAT output fields using Xpath

• Full output generated and cached per schedule, also available via API for custom queries

• Elements can be hardcoded or given default values

• No ability to translate between different domains, cascade through multiple elements, or suppress invalid values
November 30th 2014

• After frenzy of activity, the deadline was met, but the landscape was already different
• Schema 1.1 released November 1\textsuperscript{st}, requiring a new approach
• GeoPortal now supports harvesting of POD records
• Mid-November summit of FGDC and GSA to begin ISO crosswalk standardization revealed many different approaches and perspectives worth consideration
Present Day

- Participation in FGDC ISO Metadata Crosswalk development – clear need for consistency
- Developed new spreadsheet for POD 1.1 – but strongly considering handling non-geo records as dcat.json rather than ISO xml
- Received and are busy staging update to GeoPortal server to support POD 1.1
- No new mandatory fields means we will meet Feb. 1st deadline with existing CSDGM records
Harvesting Aside

• We currently harvest all our records (aside from two NGDA records) to data.gov via data.json
• CSW harvest fails because CKAN can only handle one metadata format per CSW endpoint
• WAF harvest fails because our WAF is dynamic and URLs contain “?”
• We’ve recently addressed these issues and are considering switching to harvesting geo records in native formats, but wary of validation issues
Path forward for ISO

- EPA remains committed to ISO implementation
- Strong user support for continued EME development, but significant work remains
- Must seriously consider adopting ISO profile in ArcCatalog
- FGDC engagement with Data.gov team is critical, need to ensure the expertise of the federal geospatial metadata community included in Data.gov decisions and aligned with ISO implementation
Acknowledgements

• Fearless leaders:
  – David Parrish and Ana Greene, EPA

• EDG team:
  – Suzanne Pierson, Catherine Harness, Baohong Ji, Steven Jett, Dan Spinosa, and Mettja Kuna, Innovate!, Inc.