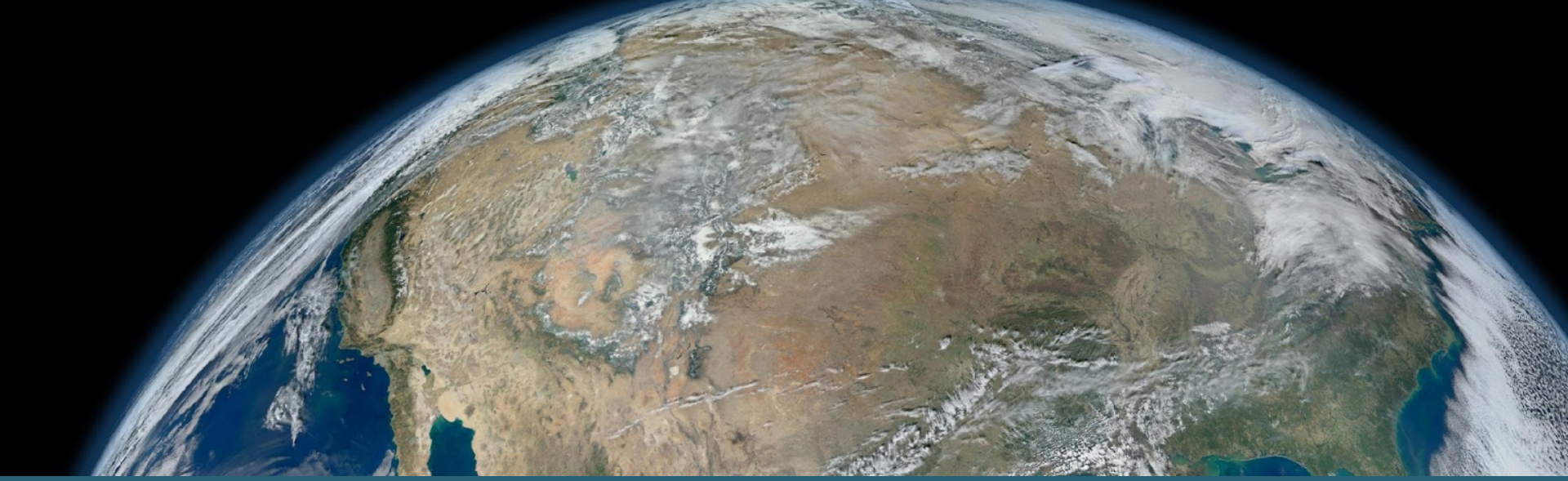


Design and implementation of automation tools for DSMM diagrams and reports

Sonny Zinn[1], John Relph[2], Ge Peng[3], Anna Milan[2], and Aaron Rosenberg[1]

Earth Resources Technology, Inc.[1]
National Centers for Environmental Information[2]
Cooperative Institute for Climate and Satellites-North Carolina[3]

January 10 & 13, 2017

NOAA Satellite and Information Service | National Centers for Environmental Information

OneStop: Data Discovery and Access

- *OneStop* supports NOAA's efforts by leveraging existing catalog and access technologies to develop an improved data access framework.
- The framework will be based on improved discovery, access, and visualization services for the data.
- One of the project activities is to provide transparent dataset quality information to users.

Data Stewardship Maturity Matrix (DSMM)

Each dataset is evaluated in 9 areas and assigned scores.

Preservability
Accessibility
Usability

Production Sustainability
Data Quality Assurance
Data Quality Control/Monitoring

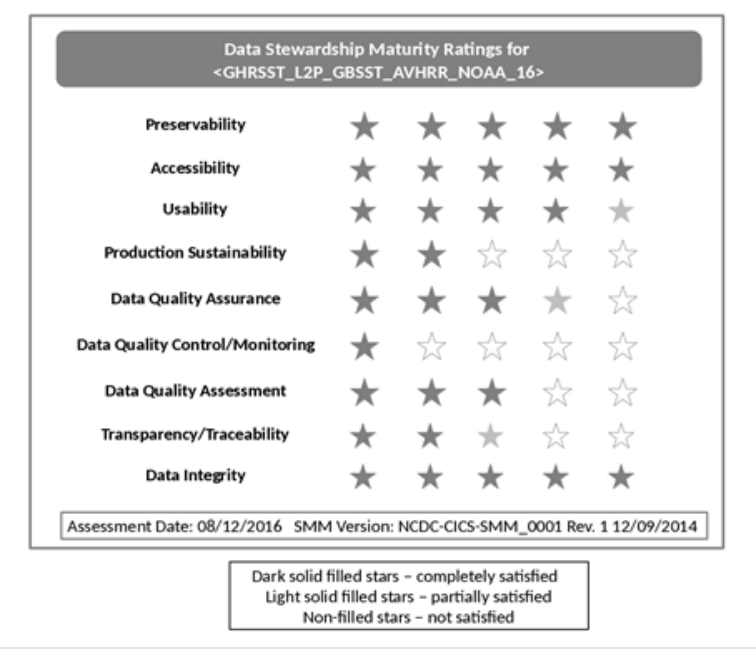
Data Quality Assessment
Transparency/Traceability
Data Integrity

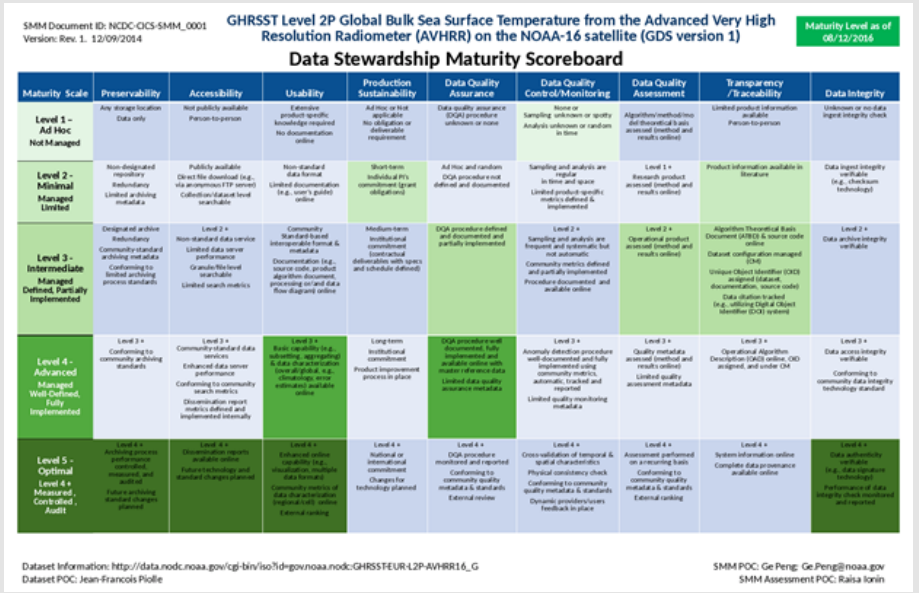
❖ **Level 5**

- Archived by NCEI, which is NOAA designated repository. NOAA is compliant to NARA standards
- Metadata following ISO 19115-2 standards.
- Compliant to OIAS RM
- Plans to update metadata to ISO 19115-1 at a later date
- Using NCEI Silver Spring Archive Management System, AMS.

DSMM Graphics - Star Rating Diagram & Scoreboard

- Great tools for summarizing DSMM.
- Used to manually generate them from a pptx template.

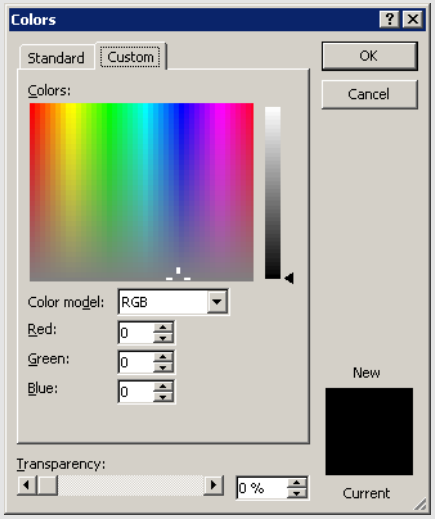




Coloring is fun but ...

- Scoreboard: 9 x 2 = 18 table cells; 7 color choices
- Star Rating Diagram: 90 places; 3 color choices

Maturity Scale	R	G	B	Color
Level 1	229	244	224	
Level 2	203	234	192	
Level 3	176	223	161	
Level 4	85	168	57	
Level 5	56	112	38	



Right Click >
Format Shape >
Shadow >
Color >
More Colors >
Custom >
Red, Green, Blue (3) >
OK
(10 clicks)

(18 + 90) x 10 = 1,080 clicks for coloring two diagrams!

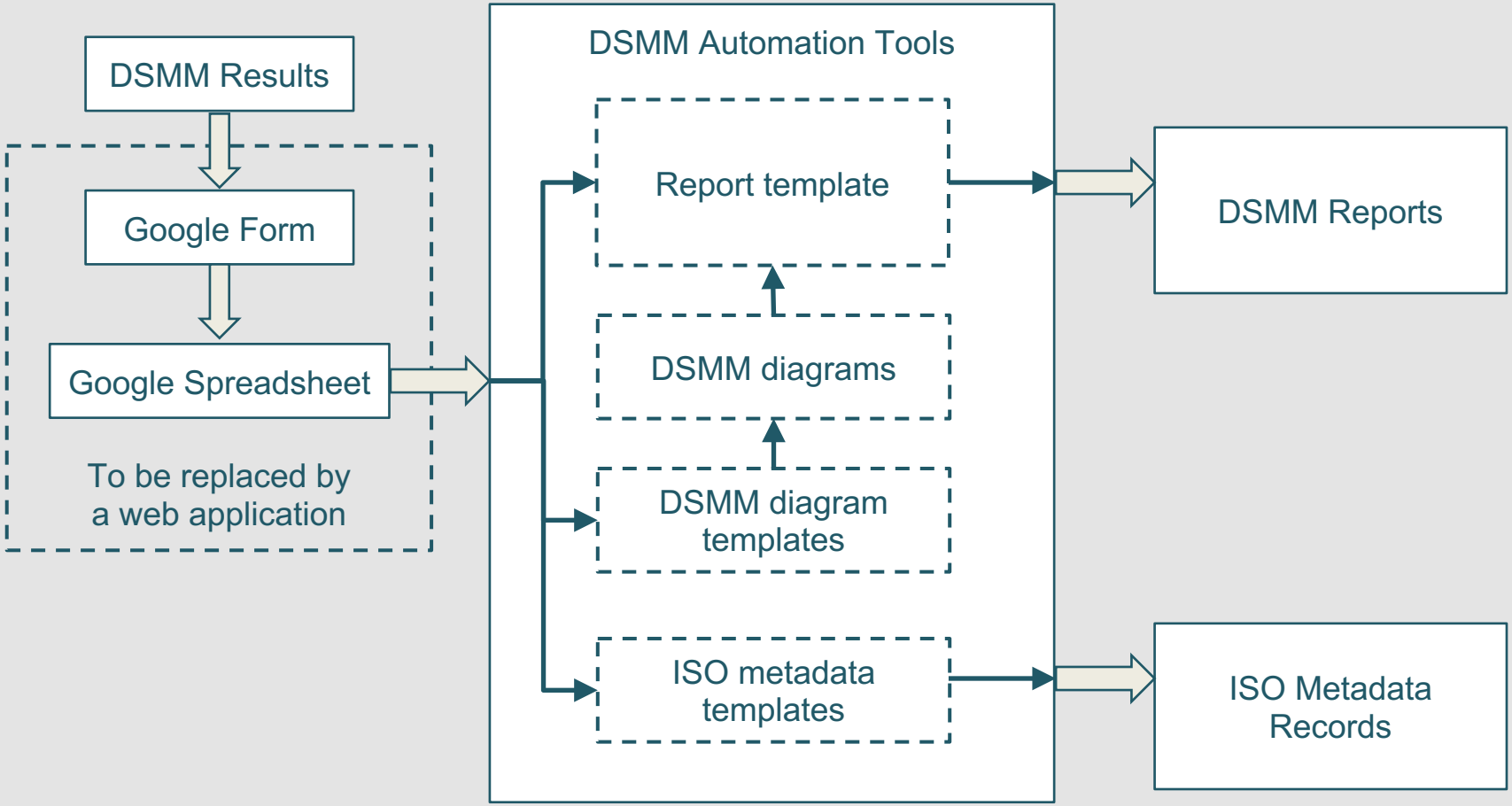
Can we do better?

- Embedded a macro inside the pptx template file.
- Users would enter 9 scores to the template and run the macro. Done!
- This was the start of the DSMM automation tools.

Table A: Maturity Scores

Preservability	5
Accessibility	4
Usability	3
Production Sustainability	2
Data Quality Assurance	1

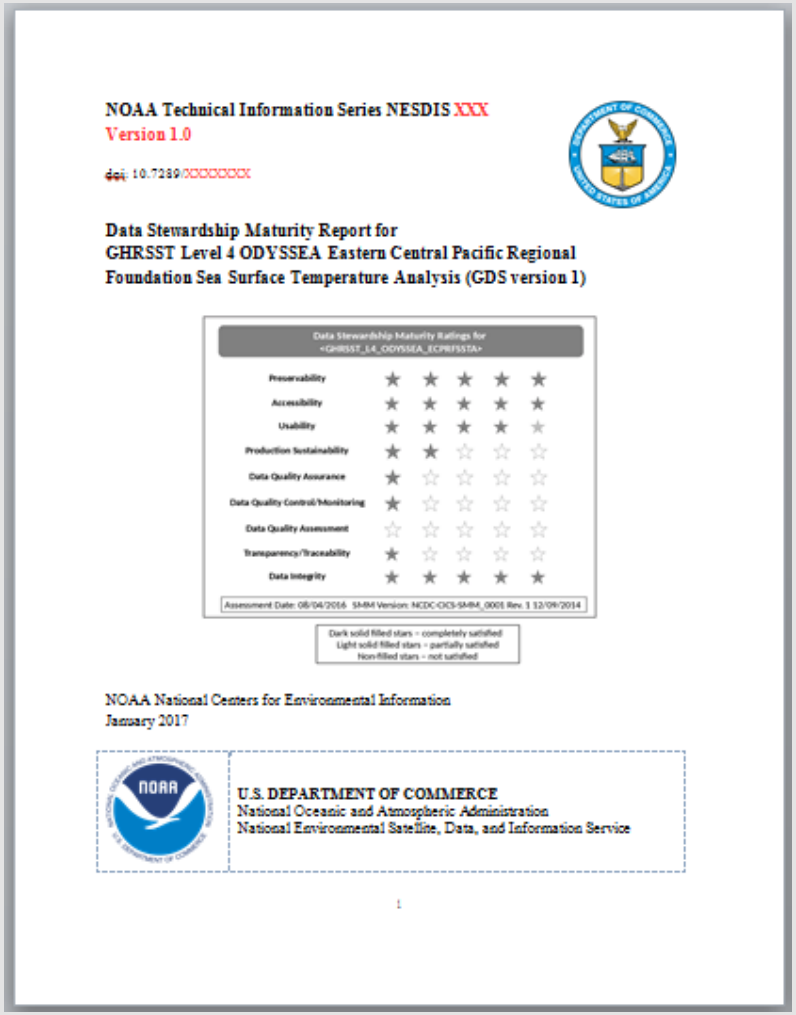
Flow of DSMM Data



Sample Report

Automated:

- Text placement
- Diagrams
- Summary of ratings
- Assessment revision history table
- Abstract if available
- Assessment tables
- References



NOAA Technical Information Series NESDIS XXX
Version 1.0

doi: 10.7289/XXXXXXXX

Data Stewardship Maturity Report for
GHRSST Level 4 ODYSSEA Eastern Central Pacific Regional Foundation Sea Surface Temperature Analysis (GDS version 1)

Data Stewardship Maturity Ratings for <GHRSST_L4_ODYSSEA_ECPRFSSTA>

Preservability	★ ★ ★ ★ ★
Accessibility	★ ★ ★ ★ ★
Usability	★ ★ ★ ★ ☆
Production Sustainability	★ ★ ☆ ☆ ☆
Data Quality Assurance	★ ☆ ☆ ☆ ☆
Data Quality Control/Monitoring	★ ☆ ☆ ☆ ☆
Data Quality Assessment	☆ ☆ ☆ ☆ ☆
Transparency/Traceability	★ ☆ ☆ ☆ ☆
Data Integrity	★ ★ ★ ★ ★

Assessment Date: 08/04/2016 SMM Version: NCDC-CICS-SMM_0001 Rev. 1 12/09/2014

Dark solid filled stars – completely satisfied
Light solid filled stars – partially satisfied
Non-filled stars – not satisfied

NOAA National Centers for Environmental Information
January 2017

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Environmental Satellite, Data, and Information Service

1

LibreOffice API for MS Office Files

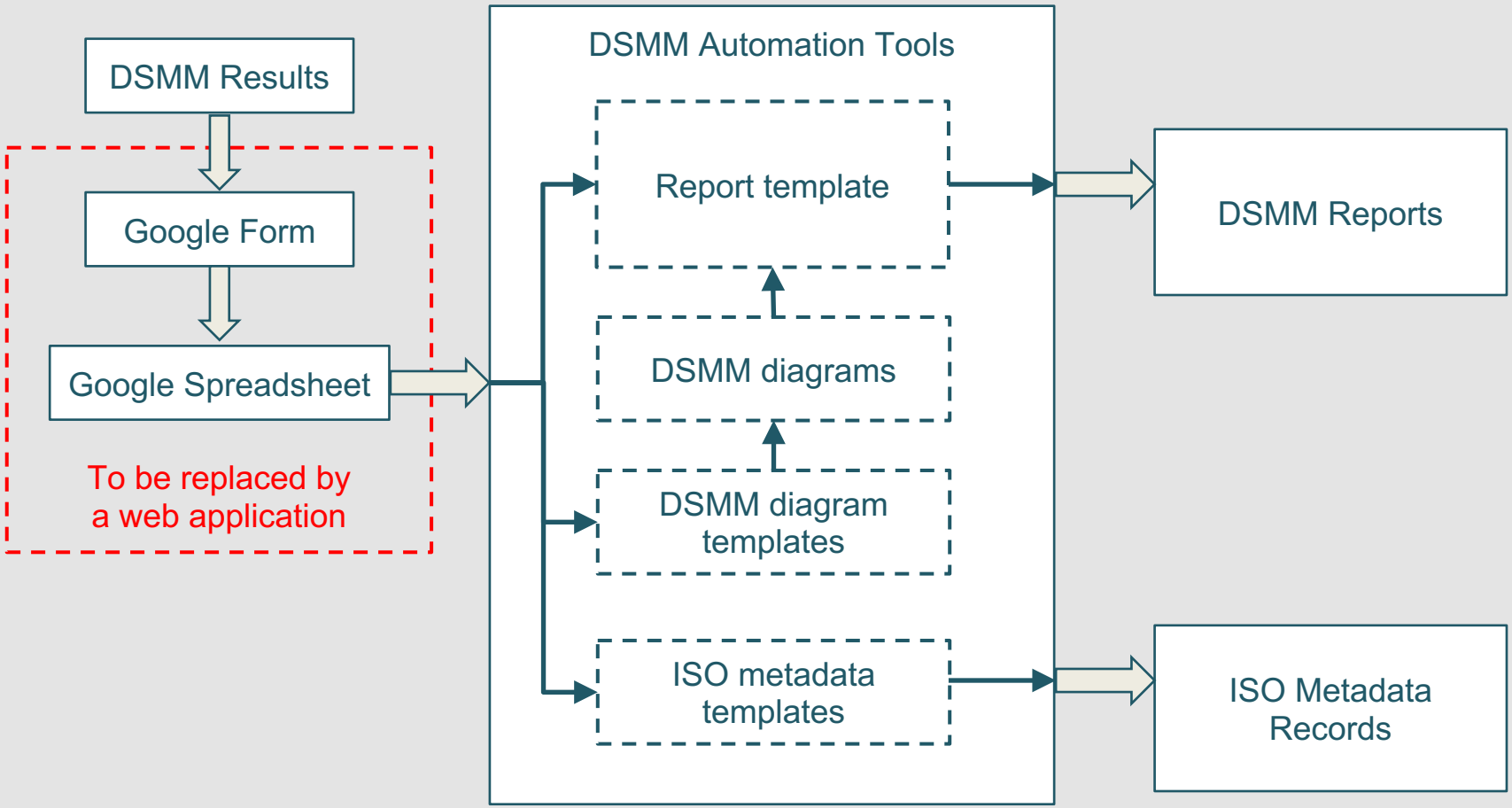
- LibreOffice Suite ~ Microsoft Office.
- LibreOffice is free and available with source code.
- Originates from StarOffice of Sun Microsystems.
- LibreOffice 4.3.7 - over 7.2 million lines of C++, Python, and Java codes.
- We use Java + LibreOffice API for report automation.

LibreOffice is Imperfect

- Issues we have worked around:
  - A page break disappeared. Fixed at XML level.
  - Table width and column widths not preserved. Fixed at XML level.
  - Image replacement at the zip file level.

- Issue still to be fixed:
  - Determining page numbers for diagrams and tables is problematic. When the template is opened with LibreOffice writer, table bodies are detached from captions and displayed on next page.

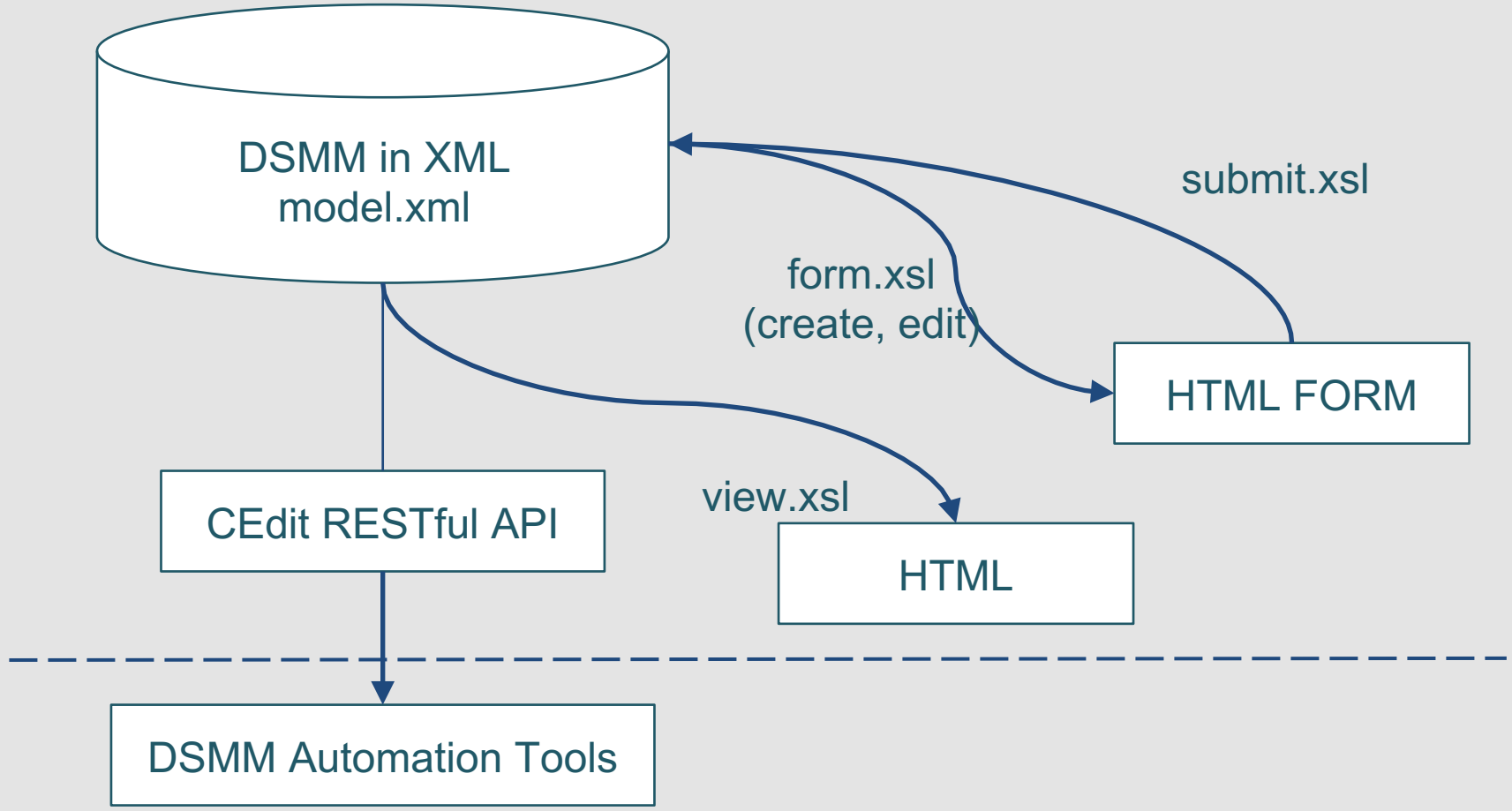
Flow of DSMM Data



Let's Use CEdit

- Google Forms - very easy to create and collect information to a spreadsheet.
- However, it is not for retrieving and editing information already entered.
- Is there an alternative? - CEdit!
  - CEdit is a metadata editor developed by NCEI.
  - Located at https://www.ngdc.noaa.gov/cedit/.
  - Stores user data in XML.
  - Requires to write XML and XSLT files.

Integration with CEdit



Conclusions & Outlook

- We built DSMM automation tools:
  - Generation of DSMM drafts to be published to NOAA Institutional Repository.
  - Production of ISO metadata records.
- We are working to complete the CEdit integration:
  - Provides a central source of DSMM results.
  - Enables metadata content editors to retrieve and edit DSMM.
- Extend CEdit to support revision history.

Acknowledgements

- IR Template
  - Robert Partee
  - Raisa Ionin
  - Paul Lemiux
  - Don Collins
- ISO Metadata Tool
  - Jason Shapiro
- DSMM Automation Tools
  - Thomas Jaensch
- CEdit
  - Rich Fozzard
  - Marty Aubrey

NCEI

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

www.ncei.noaa.gov

NCEI Climate Facebook: http://www.facebook.com/NOAANCEIclimate
NCEI Ocean & Geophysics Facebook: http://www.facebook.com/NOAANCEIoceangeo
NCEI Climate Twitter (@NOAANCEIclimate): http://www.twitter.com/NOAANCEIclimate
NCEI Ocean & Geophysics Twitter (@NOAANCEIocngeo): http://www.twitter.com/NOAANCEIocngeo

Backup Slides

For more information...

Structure of DOCX File

- A zip file contains more files.
- Contents:
  - word/document.xml
  - Embedded images:
  - word/media/image1.png
  - word/media/image2.png
  - etc.

```
Name
----
[Content_Types].xml
_rels/.rels
word/_rels/document.xml.rels
word/document.xml
word/footer1.xml
word/footnotes.xml
word/endnotes.xml
word/media/image4.png
word/theme/theme1.xml
word/media/image3.png
word/media/image1.png
word/media/image2.png
word/settings.xml
word/numbering.xml
word/styles.xml
docProps/app.xml
customXml/_rels/item1.xml.rels
customXml/itemProps1.xml
customXml/item1.xml
word/fontTable.xml
word/webSettings.xml
word/stylesWithEffects.xml
docProps/core.xml
-------
23 files
```

Details on DSMM Report Generation

- We used placeholders like {DSMM_DATASET_SHORT_NAME} throughout the template to mark where to place a text string.

- Tables do not need a placeholder. They are structured and easy to identify.

Table 2. Stewardship Maturity Levels and Detailed Justifications for Each of Nine DSMM Key Components for the <{DSMM_DATASET_SHORT_NAME}> Dataset.

DSMM Key Component	**Stewardship Maturity Rating, Justification, and Comments**
Preservability	
Accessibility	

Problem Directly Working with XML

- Suppose we want to search and replace text.
- Could walk through XML nodes to find the target text.

Paragraph <w:p>
Run <w:r>
Text <w:t>

- BUT what if our search string crosses a run boundary?

Paragraph May Not Be a Single Piece

The information about dataset and stewardship maturity assessment is summarized in Table 1. The data stewardship maturity ratings are displayed as the scoreboard (Figure 1) and rating diagram (Figure 2) with the detailed justifications in Table 2.

```
3787            <w:r>
3788                <w:rPr>
3793                <w:t>The information about dataset and stewardship maturity assessment is summarized in Table 1. The data stewardship mat
3793 (Figure 1)</w:t>
3794            </w:r>
3795            <w:r>
3796                <w:rPr>
3802                <w:t xml:space="preserve"> </w:t>
3803            </w:r>
3804            <w:r>
3805                <w:rPr>
3810                <w:t>and rating diagram (Figure 2) with the detailed justifications in Table 2.</w:t>
3811            </w:r>
3812            <w:r/>
3813        </w:p>
3814        <w:p>
```