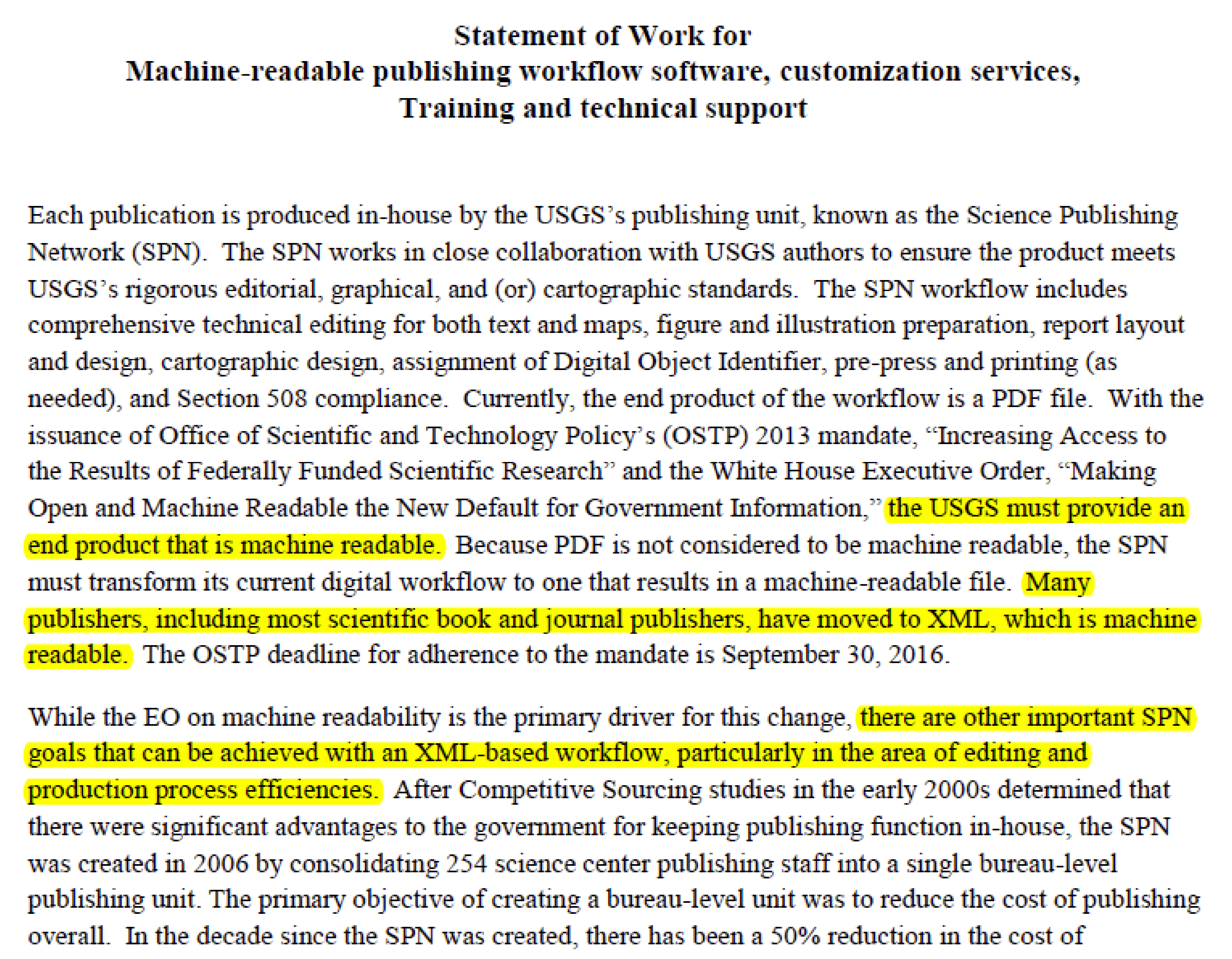
INERA
eXtyles | edifix

eXtyles®

Editorial and XML
Solutions for
Publishers

Why Are We Here?



Statement of Work for
Machine-readable publishing workflow software, customization services,
Training and technical support

Each publication is produced in-house by the USGS's publishing unit, known as the Science Publishing Network (SPN). The SPN works in close collaboration with USGS authors to ensure the product meets USGS's rigorous editorial, graphical, and (or) cartographic standards. The SPN workflow includes comprehensive technical editing for both text and maps, figure and illustration preparation, report layout and design, cartographic design, assignment of Digital Object Identifier, pre-press and printing (as needed), and Section 508 compliance. Currently, the end product of the workflow is a PDF file. With the issuance of Office of Scientific and Technology Policy's (OSTP) 2013 mandate, "Increasing Access to the Results of Federally Funded Scientific Research" and the White House Executive Order, "Making Open and Machine Readable the New Default for Government Information," the USGS must provide an end product that is machine readable. Because PDF is not considered to be machine readable, the SPN must transform its current digital workflow to one that results in a machine-readable file. Many publishers, including most scientific book and journal publishers, have moved to XML, which is machine readable. The OSTP deadline for adherence to the mandate is September 30, 2016.

While the EO on machine readability is the primary driver for this change, there are other important SPN goals that can be achieved with an XML-based workflow, particularly in the area of editing and production process efficiencies. After Competitive Sourcing studies in the early 2000s determined that there were significant advantages to the government for keeping publishing function in-house, the SPN was created in 2006 by consolidating 254 science center publishing staff into a single bureau-level publishing unit. The primary objective of creating a bureau-level unit was to reduce the cost of publishing overall. In the decade since the SPN was created, there has been a 50% reduction in the cost of

USGS Solicitation: Key Requirements

- Machine-readable end product
- XML
- Editorial and production process efficiencies

What Is Machine-Readable Content?

Formats should be machine-readable (i.e., data are reasonably structured to allow automated processing). Open data structures do not discriminate against any person or group of persons and should be made available to the widest range of users for the widest range of purposes, often by providing the data in multiple formats for consumption. To the extent permitted by law, these formats should be non-proprietary, publicly available, and no restrictions should be placed upon their use.

M-13-13 — Memorandum for the Heads of Executive Departments and Agencies
Open Data Policy—Managing Information as an Asset
(from project-open-data.cio.gov)

An important starting point is to understand that “machine readable” is not synonymous with “digitally accessible” information. Scanning a report, the text, graphics, or even rows and columns of numbers, makes it digitally accessible, but a computer still is not really able to “understand” the information. This distinction can be seen in the difference between a magazine cover and a barcode on that cover. A computer cannot directly understand what the picture on the magazine represents, even if it is presented in an online format, but it can read and understand the bar code, using it for identifying the price and tracking the purchase, for example.

A Primer on Machine Readability for Online Documents and Data
(from data.gov)

Why Isn’t PDF “Machine Readable”?

- Used to be a proprietary format
  - Became an open standard in 2008
  - Still includes some proprietary technologies that can limit accessibility
- Back to that magazine cover example…
  - PDF is a visually formatted representation of a document
  - Richer than it used to be, now with selectable text, active links, alt-text for accessible images
  - Still lacks metadata and structured text

Enter XML

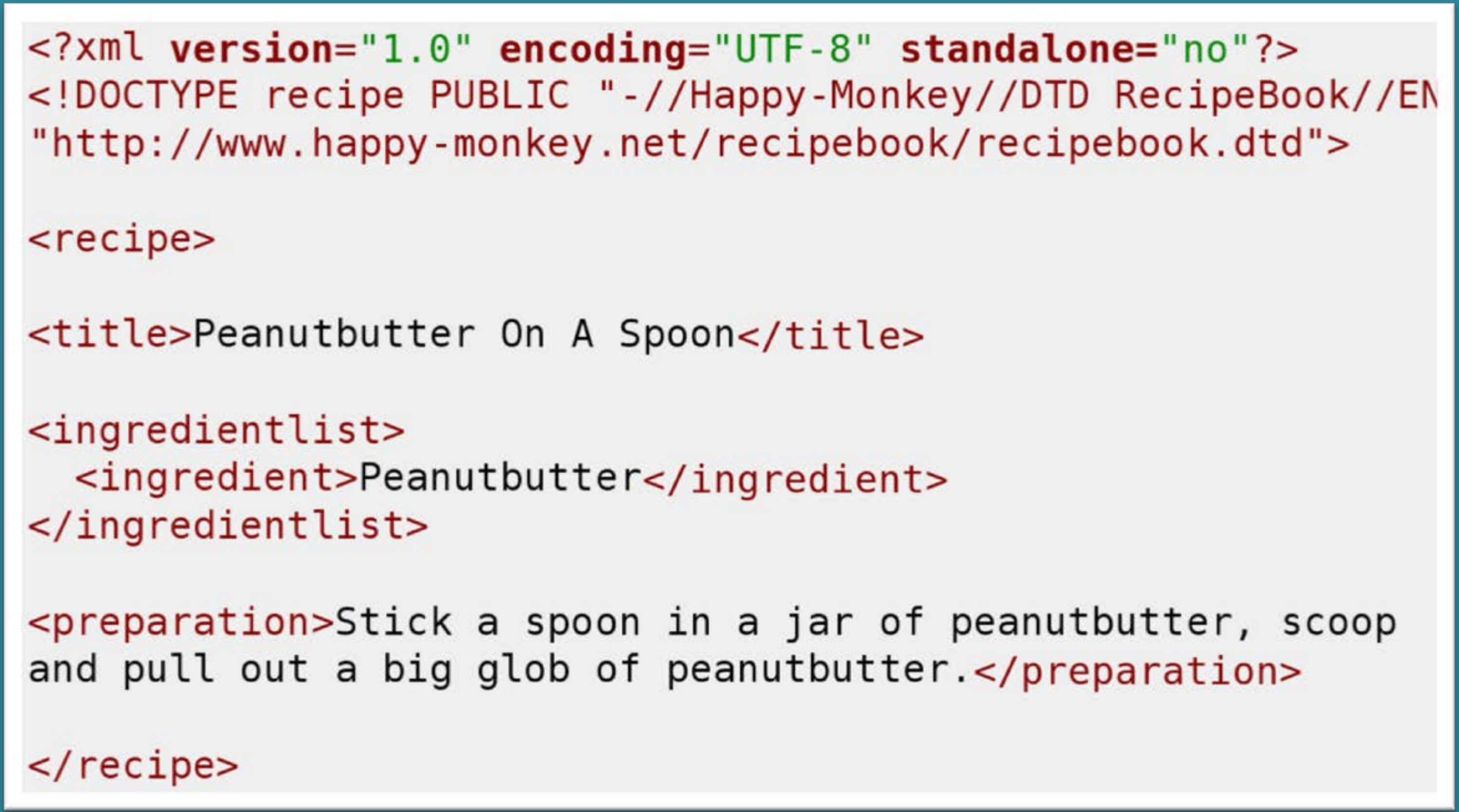
eXtensible Markup Language

- A markup language that defines a set of rules for encoding content in a format that is both **human readable and machine readable**
- Wait… is XML really *human* readable?

Use XML, they said.
It will be fun, they said.

XML 101

Or, How I Learned to Stop Worrying and Love Angle Brackets



```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE recipe PUBLIC "-//Happy-Monkey//DTD RecipeBook//EN
"http://www.happy-monkey.net/recipebook/recipebook.dtd">

<recipe>

<title>Peanutbutter On A Spoon</title>

<ingredientlist>
  <ingredient>Peanutbutter</ingredient>
</ingredientlist>

<preparation>Stick a spoon in a jar of peanutbutter, scoop
and pull out a big glob of peanutbutter.</preparation>

</recipe>
```

The eXtyles Pilot Phase

- For the past six months, your colleagues have tested eXtyles for USGS
- They provided Inera valuable feedback to ensure that eXtyles is appropriately customized
- As a result, eXtyles has been road-tested at USGS on actual USGS content

Do I Really Need to Know This?

- All eXtyles operations take place in Microsoft Word, on Word documents (.doc or .docx)
- eXtyles collects metadata and applies document structuring in Word so an XML file can be produced at the end of the process with a single mouse click
- Having said that…

INERA
eXtyles | edifix

Just Enough to Be Dangerous

- USGS documents are long and complex
- You *will* at times get errors when exporting XML from Word
- You don't need to be an XML expert, but it helps to have a basic understanding of a few key XML principles

Key XML Terms

- ## Tag

  A markup construct that semantically defines the content. Tags typically surround text and start and end with angle brackets (< and />).

  A tag can also be "empty," i.e., not surrounding text:

  <line-break/>

- ## Element

  A logical document component that has start and end tags:

  <book-title>Restoration Handbook for Sagebrush Steppe</book-title>

More Key XML Terms

- ## Attribute

  A markup construct consisting of element + value, set in quotation marks:

  <contrib contrib-type="author">

- ## DTD

  Document Type Definition: defines the document structure with a list of legal **elements** and **attributes**; it is the ruleset for your XML. USGS is using the Book Interchange Tag Suite aka **BITS DTD**. (*Peanutbutter on a Spoon* uses the RecipeBook DTD.)

- ## Valid

  Valid XML conforms to the rules of a DTD; invalid XML does not.

INERA
eXtyles | edifix

It’s All Semantics

*XML tags identify what a document component **is**. They do not identify what a document component **looks like**.*

- XML tells us that this is a title:

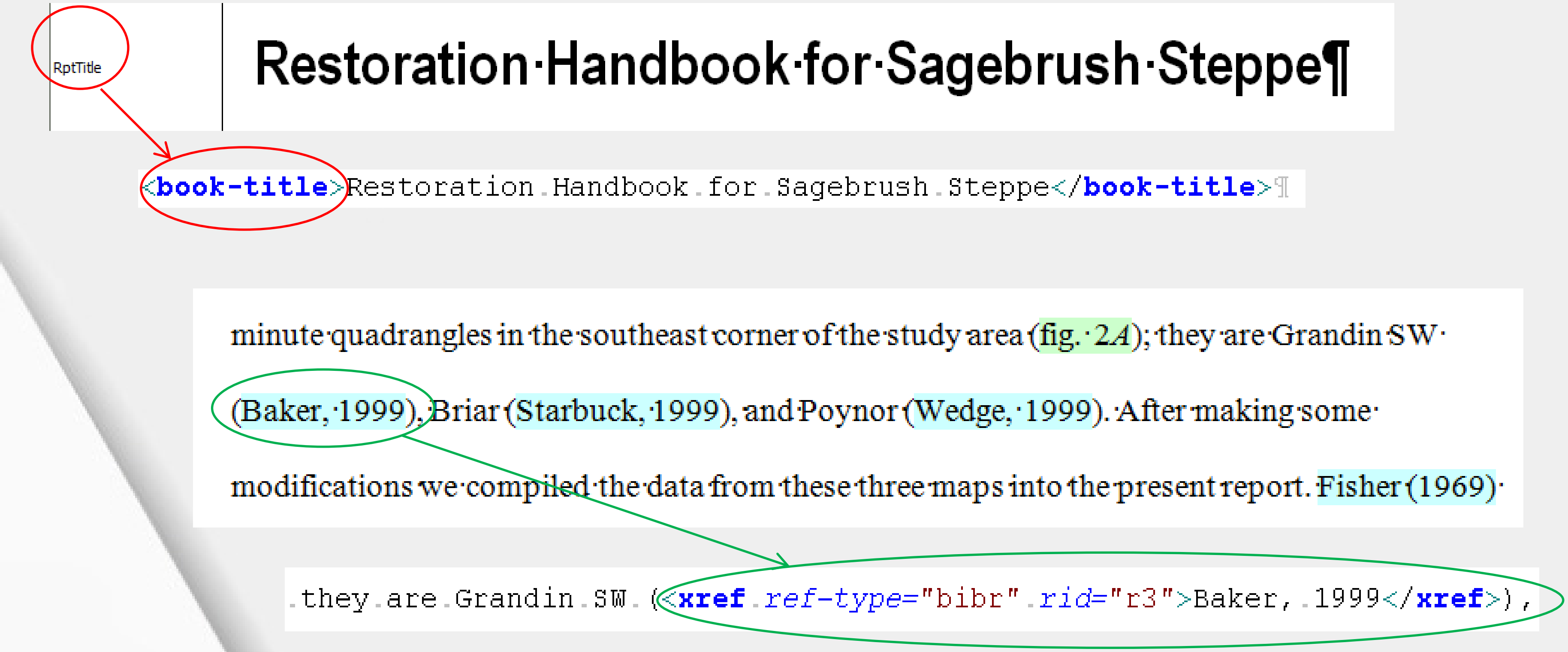
```
<book-title>Restoration Handbook for
   Sagebrush Steppe</book-title>
```

- XML does not tell us what this title should look like (e.g., font size, face markup, etc.):

Restoration Handbook for Sagebrush Steppe

eXtyles and XML

eXtyles maps Word paragraph and character styles to specific tags in the XML



Restoration·Handbook·for·Sagebrush·Steppe¶

```
<book-title>Restoration.Handbook.for.Sagebrush.Steppe</book-title>¶
```

minute·quadrangles·in·the·southeast·corner·of·the·study·area·(fig.·2*A*);·they·are·Grandin·SW·(Baker,·1999),·Briar·(Starbuck,·1999),·and·Poynor·(Wedge,·1999).·After·making·some·modifications·we·compiled·the·data·from·these·three·maps·into·the·present·report.·Fisher·(1969)·

```
.they.are.Grandin.SW.(<xref.ref-type="bibr".rid="r3">Baker,.1999</xref>),
```

Benefits of XML for USGS Publications

- Machine-readable content
  - XML can be transformed into online content that meets accessibility standards
  - XML is a non-proprietary, human-readable format for long-term archiving of electronic content
- Multi-format publication
  - XML can be used for PDF and HTML production
  - XML can also be easily transformed into formats such as ePub and Mobi, which enable content to be read on devices such as Kindle and Nook

Benefits of XML for USGS Editors

Structured Word documents enable context-sensitive automated editing

- eXtyles was designed to apply to Word documents the structure required to produce valid XML from Word
- This document structuring can also be leveraged to enforce editorial style and check document accuracy and completeness

Some Key eXtyles Editorial Tools

- **Paragraph styles**: eXtyles can apply editorial rules to certain document elements, such as deleting punctuation at the ends of headings or enforcing a preferred format for figure titles
- **Bibliographic references:** eXtyles can automatically copyedit journal and book references according to USGS's editorial style preferences
- **In-text citations/callouts:** eXtyles can tell you if a reference or figure is uncited, or if a callout to a table is out of order

INERA
eXtyles | edifix

So… What's My One Big Takeaway from All This?

- **Accurately applied paragraph styles**
- The eXtyles paragraph style template resembles, but does not duplicate, previous USGS paragraph style templates
- An XML workflow is less forgiving of misapplied styles than a design-focused workflow
- We will provide thorough documentation and guidance

INERA
eXtyles | edifix