Federal Sitemaps: An XML-Based Standard for Searching the Invisible Web

Presentation at the XML CoP Meeting
Mills Davis and Brand Niemann, SICoP Co-Chairs, and
JL Needham, Google
January 17, 2007
Part of Building DRM 3.0 and Web 3.0 by Managing Context Across Multiple Documents and Organizations

Overview

- 1. History and Wiki Page
- 2. EPA Experience
- 3. Proposed Pilot
- 4. Schedule
- 5. Questions and Answers

1. History and Wiki Page

- The Sitemap protocol is an open, XML-based standard for managing search engine crawling. The protocol provides website owners a means of communicating to search engines the location, priority, change frequency, and last modification date of all pages on a website or web-accessible database, which can ensure complete and efficient crawling of the site's contents.
- The Sitemap protocol was introduced by Google in June 2005 under a Creative Commons License and was adopted in November 2006 as an industry standard by Google, Microsoft and Yahoo.
  - See SearchEngineWatch - Search Engines Unite On Unified Sitemaps System, November 16, 2006.
- FederalSitemaps is an initiative to help federal agencies make their websites more accessible to search engine users through sitemapping.
  - See recent presentation to OMB and SICoP.

See http://colab.cim3.net/cgi-bin/wiki.pl?FederalSitemaps

2. EPA Experience

- Sitemaps augments, but does not replace regular crawling.
- Sitemaps is focused on exposing the contents of databases which estimates suggest may be as much as 90% of Web content.
- The current Sitemaps protocol is the "lowest-common-denominator" approach (see next slide)
- In EPA's new template, we're including the Dublin Core fields that make us consistent with the eGov Act of 2002 and the OMB guidance pursuant to it (see slide 6).
- I will meet with the Searchmasters and discuss how we might alter our existing "jump pages" to conform to the Sitemap protocol, or to alter our jump-page creation process to also create Sitemaps.

Source: John Shirey, Notes on Federal Sitemaps Discussion, January 10-11, 2007.

2. EPA Experience

- <?xml version="1.0" encoding="UTF-8"?><urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"> <url> <loc >http://www.example.com/</loc> *<lastmod >2005-01-01</lastmod>* *<changefreq >monthly</changefreq>* *<priority >0.8</priority>* </url></urlset>

2. EPA Experience

- <html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en"><!--EPA Template version 3.2.1, 28 June 2006 --><head> <title>Page Title | Area Name | US EPA</title> <meta name="DC.title" content="" /> <meta name="DC.description" content="" /> <meta name="keywords" content="" /> <meta name="DC.Subject" content="" /> <meta name="DC.type" content="" /> <!-- For date metadata, use the format YYYY-MM-DD --> <meta name="DC.date.modified" content="" /> <meta name="DC.date.created" content="" /> <meta name="DC.date.reviewed" content="" /> <meta name="DC.language" content="en" /> <meta name="DC.creator" content="" /> <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" /> <link rel="meta" href="http://www.epa.gov/labels.rdf" type="application/rdf+xml" title="ICRA labels" /> <meta http-equiv="Content-Style-Type" content="text/css" /> <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />

Source: John Shirey, New EPA Basic Template, January 8, 2007.

2. EPA Experience

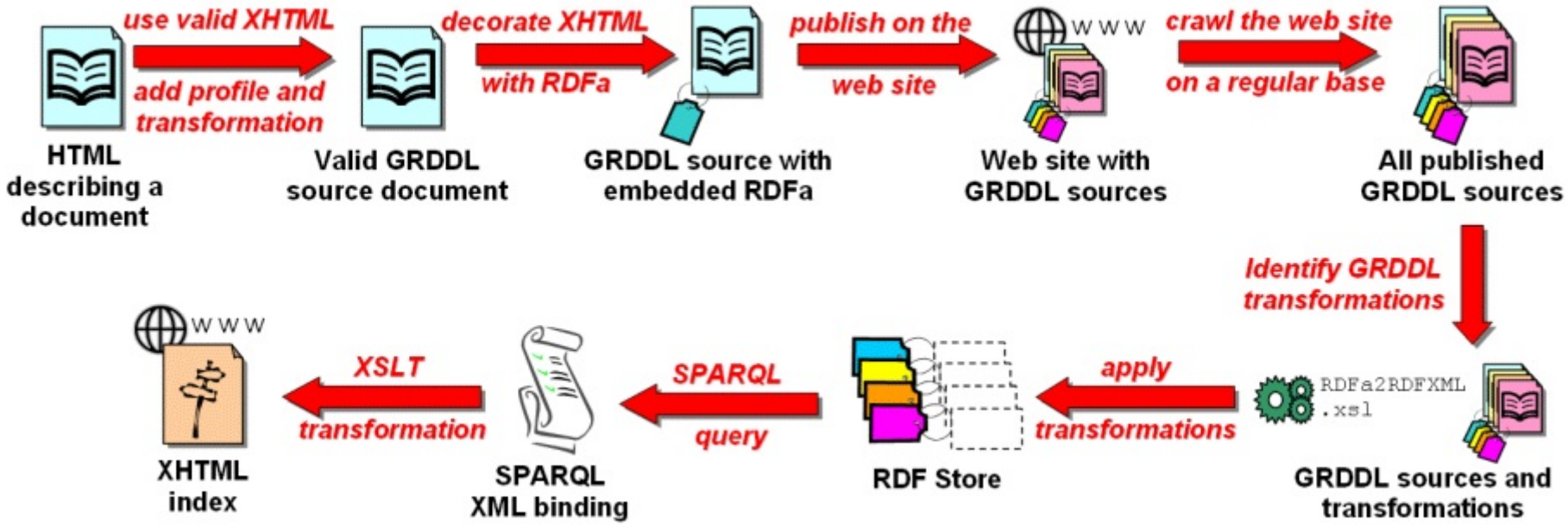
- “Sitemaps as a method for discovering database content is something that I heartily endorse. It makes sense, and it's good to have a data standard for doing it. Google, et. Al. are to be commended for that. Too bad it's such a minimalist protocol! As we work to expose database contents to our internal search engine, we will keep in mind the need to express that content in a Sitemap protocol as well. EIMS is our first target database, hopefully tackling it this spring.”

Source: John Shirey, Notes on Federal Sitemaps Discussion, January 10, 2007.

3. Proposed Pilots

- Pilot Outline:
  - Tools, Collaboration, and Content (details to be added by Mills Davis) based on Kapow Technical Roundtable, January 9, 2007, and other meetings.
- Example: Gleaning RDF from XML (GRDL) as at relates to the Google Sitemaps, Semantic Wikis and X-Forms Use Cases (see next slide):
  - http://www-sop.inria.fr/acacia/personnel/Fabien.Gandon/tmp/grddl/rdfaprimer/PrimerRDFaSection.html

3. Proposed Pilots



In this example the focus is on automating the construction of indexes. The idea is to crawl GRDDL source documents and extract embedded RDFa to feed an RDF store. SPARQL queries are then solved against this store and rendered as web pages to automatically generate up-to-date indexes.

4. Schedule

- **January 17, 2007, XML CoP**:
  - About 15 minutes to introduce the protocol to ~25 XML experts and advocates across federal agencies and the effort we're undertaking to encourage its adoption. May discuss the white paper idea and the prospective conference on the protocol in coming months.
- **January 29, 2007, EPA**:
  - One hour presentation to EPA web managers on how to implement the protocol to open EPA sites now closed to search engine crawlers. This is an opportunity to observe how we approach discussions with a major federal agency.
- **February 15, 2007, Web Content Managers Forum (tentative)**:
  - Conference call involving 100-200 Forum participants in which Google SICoP and representatives of NCES, OSTI, and PlainLanguage.gov will discuss in detail various approaches to opening flat file, fielded and other dynamic databases to crawling with the protocol.
- **March 20-22, 2007, FOSE 2007**:
  - Possible panel slot for an introduction of the protocol in the FIRM Forum at FOSE. Also requesting three one-hour tutorial sessions like last year on implementing DRM 2.0.
- **April-May 2007, Federal Sitemaps Conference/Workshop**:
  - A conference dedicated to the protocol based on further discussions of the audience it would target and who would contribute. Vint Cerf, Google CTO, who is following progress of this effort, would probably keynote.

4. Schedule

- January 29th Meeting at EPA on Building DRM 3.0 and Web 3.0:
  - Sitemaps:
    - JL Needham, Google
  - Digital Library (see next slide):
    - JL Needham, Google
  - SICoP Special Conference, February 6, 2007, and Pilot (see slide 13):
    - Mills Davis, SICoP Co-Chair
  - General Discussion:
    - Brand Niemann, EPA Enterprise Architecture Team

4. Schedule

- Federal Sitemaps:
  - http://colab.cim3.net/cgi-bin/wiki.pl?FederalSitemaps
- Digital Library:
  - Coincidental to our recent communication about the Sitemap protocol and how it can make EPA and federal agency information more accessible to citizen users, my colleagues in Google's Library Partnerships team have approached me about reopening discussions with EPA about the prospect of Google digitizing a segment of the EPA library's print holdings and making them accessible through Google's services including Google Book Search. As a former member of this Library Partnerships team, I know we held discussions with the director of EPA's Information Access Division about this prospect in early 2006. Then, as now, the offer was that Google would conduct this digitization at no expense to EPA.

4. Schedule

- SICoP Special Conference, February 6, 2007, and Pilot:
  - Building DRM 3.0 and Web 3.0 by Managing Context Across Multiple Documents and Organizations:
    - http://colab.cim3.net/cgi-bin/wiki.pl?SICoPSpecialConference_2007_02_06
  - Pilot:
    - Tools, Collaboration, and Content (details to be added by Mills Davis) based on Kapow Technical Roundtable, January 9, 2007, and other meetings.

5. Questions and Answers

- John Lewis (JL) Needham
  - Strategic Partner Development Manager, Google, Inc.
  - jlneedham@google.com
- Mills Davis
  - Project10x and SICoP Co-Chair
  - mdavis@project10x.com
- Brand Niemann
  - EPA Enterprise Architecture Team and SICoP Co-Chair
  - niemann.brand@epa.gov