METS and MODS / MINERVA
PART 3
METS Profiles for Web Sites

- Leslie D. Myrick, NYU

DLF Forum, Spring 2004

Today’s Topics

- Background: challenges of web archiving
- How METS can resolve some of the challenges
- Why METS Profiles are necessary
- Announcement of imminent posting of profiles for web sites by LC and PCWA
- “Superclass” / “subclass” modeling of those profiles vis a vis each other

Political Web Communications Archiving Project

- Under auspices of CRL and Mellon
- Participants: Cornell University, Stanford University, UT Austin, NYU
- Focus: SE Asia, Sub-Saharan Africa, Latin America, Western Europe
- Content: Internet Archive
- Mentoring: LC MINERVA / NDMSO

Web Archiving Challenges: I

- Definition of the object “web site” and its boundaries
  - what to do with external links? near files?
- Complex nature of web site structure
- Complex nature of a web page itself
  - HTML wrapper around embedded files
  - perhaps dynamically generated

Web Archiving Challenges: 2

Version Control-related storage and access issues

- Creator-driven changes: successive harvests and versions
- Repository-driven changes: refreshing, migration, other changes

Web Archiving Challenges 3

- Lack of influence over the production of the content we're archiving
- Bad metadata from producers of web pages
  - Programmatically extracted metadata as possible source for MODS and MIX bits in METS instances for websites
  - Descriptive MD from <title>, <meta> tags ??
  - Technical MD from embedded file information

http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/sdm.xml - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/sdm.xml Go Links

Back to Cross-Collection Search

Title: Democratic Socialist Movement
Date Captured: 20030417
Abstract: Socialist news, policies and Marxist analysis, socialist campaigns, anti-war campaigns, support for workers' struggles

Links

[Back to Homepage]

Democratic Socialist Movement: TIME FOR REAL CHANGE
Democratic Socialist Movement
Democratic Socialist Movement: TIME FOR REAL CHANGE
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement
Democratic Socialist Movement

Search for

Search Clear

CRL Political Web Archiving Project METS Viewer

http://dlibdev.nyu.edu/webarchive/metstest/www.socialistnigeria.org/index.html

Democratic Socialist Movemen

For struggle, Solidarity and Socialism in Nigeria

Join DSM
Contact DSM
About us
Our Manifesto
Statements

Socialist Democracy
Newspaper of the DSM

Campaigns
NCP
Trade Unions
Students
Women

2003 Elections

16 April 2003

THE APRIL 12 NATIONAL ELECTION WAS ANYTHING BUT FREE AND FAIR

Statement To Press Conference Organised By The Lagos State Chapter Of National Conscience Party (NCP) On Wednesday 16th April, 2003 At The Party Secretariat, 36a, Acme Road, Ogba-Ikeja, Lagos

SEGUN SANGO
Chairman, Lagos State NCP
(Segun Sango is also General Secretary of the Democratic Socialist Movement)

Open in New Window Open in Full Frame

start Exceed emacs@dlibdev.... 2 Internet Expl... Eudora - [In] dlibdev.nyu.edu ... Microsoft PowerP... 2:42 PM

The Case of the Purloined Metadata

The Case of the Purloined Metadata, continued

```
<snip>
<HTML>
<!-- saved from url=(0041)http://www.sport.de/spart/sk1/ski006.php3 -->
<HEAD>
<TITLE>Bienvenue sur le site de Front Social</TITLE>
<META CONTENT="text/html; charset=windows-1252" HTTP-EQUIV="Content-Type">
<META CONTENT="Sport sports Baseball Basketball Beach-Volleyball Bob Boxen Bundesliga
Bundesligavereine Championsleague DEL DFB DFB-Pokal Eishockey Ergebnisse Europameisterschaft
Europapokal Fernsehen Football Formel1 Formel3 Fußball Golf Hallenmasters Handball Hockey Inline-Skating
Leichtathletik Motorbike Motorrad Motorsport Nationalmannschaft NBA NFL NHL Reiten Rodeln Schwimmen
Skifahren Skispringen Snowboard Sportarten Sportnachrichten Surfen Tennis Tischtennis Turniere Uefa-Cup
US Open Vereine Volleyball Wassersport WBA WBC WBO Weltmeisterschaft Weltrangliste Wimbledon Fußball
Motorsport Radsport Volleyball Sport Eishockey Skisport Boxen Handball Leichtathletik Pferdesport
Schwimmen" NAME="keywords">
  <META CONTENT="Sport Sportnachrichten Sportvereine Ergebnisse Tabellen Ranglisten Bundesliga DEL
Formel 1 Tennis" NAME="description">
  <META CONTENT="thu, 30 mar 2000 12:00:00 GMT" HTTP-EQUIV="date">
  <SCRIPT language="JavaScript" SRC="sport_fichiers/sidiscript.js">
  <SCRIPT language="JavaScript">
<!--
var on = "/ima/pfeil_weiss2.gif";
var off = "/ima/pfeil_weiss.gif";
</snip>
```

Whence Web Archive Metadata?

- No dearth of metadata provided by crawlers as humble as wget or as sophisticated as Alexa
- Management of it all: a perfect job for METS
- Typical Alexa / IA SIP = .arc and .dat files along with byte offset .ndx files
  - IA .arc = 100 MB gz archive file packed with archived files from web crawl along with server's http response headers for each file.

Typical IA .arc snippet

```
<snip>
[ crawler's file header]
http://www.apgawomen.org:80/calender.htm 63.241.136.203 20030417223125
text/html 2570

[http headers]
HTTP/1.1 200 OK
Date: Thu, 17 Apr 2003 21:35:43 GMT
Server: Apache/1.3.27 (Unix) FrontPage/5.0.2.2510
Last-Modified: Sun, 26 Jan 2003 04:05:37 GMT
ETag: "3b01d2-8fb-3e335e91"
Accept-Ranges: bytes
Content-Length: 2299
Connection: close
Content-Type: text/html

[file itself]
<html>
<head>
<title>calender</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>
<body bgcolor="#FFFFFF">
</snip>
```

Technical Metadata Sources

- Crawler frontier application
- Host Server (HTTP response headers)
- Captured Files
  - Some info from accompanying HTTP headers
  - Some from post-processing with ImageMagick and its ilk for other multimedia

Technical Metadata Desiderata

- Host Server
  - IP Address, operating system, webserver configuration
- Capture Transaction
  - Timestamp, software, HTTP response headers, errors
- All Captured Files
  - File format, file name, file size, last modified date, creating software, creating hardware, generated checksum
- HTML Pages
  - Charset, language, encoding, links broken down by type, embedded scripting, <meta> tags.

Image Files

(After post-processing with ImageMagick)

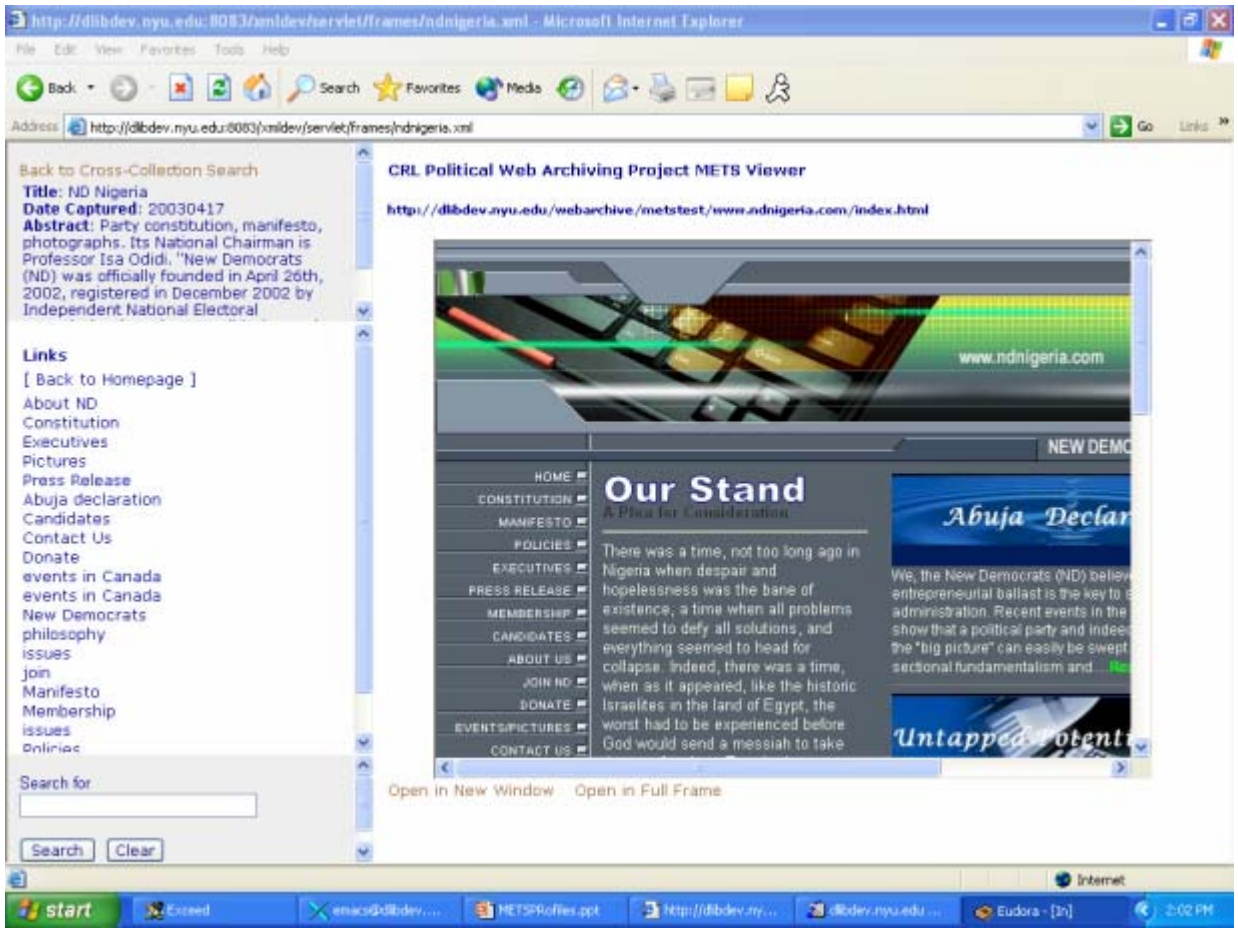
- format
- bit depth
- compression
- resolution
- imageWidth and imageLength
- file size
- identifier

Image Magick dump for Mao1925.jpg

Image: Mao1925.jpg
Format: JPEG (Joint Photographic Experts Group JFIF format)
Geometry: 142x185
Class: DirectClass
Type: true color
Depth: 8 bits-per-pixel component
Colors: 11423
Resolution: 300x300 pixels
Filesize: 8115b
Interlace: Plane
Background Color: grey100
Border Color: #DFDFDF
Matte Color: grey74
Iterations: 0
Compression: JPEG
signature:
8c173bd33c3e5667d27e51aee539afcd58ccbc8d4a11ab76b127408905f598fd
Tainted: False


```
<mix:mix>
    <mix:BasicImageParameters>
        <mix:Format>
            <mix:MIMEType>image/jpeg</mix:MIMEType>
            <mix:ByteOrder>little-endian</mix:ByteOrder>
            <mix:Compression>
                <mix:CompressionScheme>1</mix:CompressionScheme>
                <mix:CompressionLevel>0</mix:CompressionLevel>
            </mix:Compression>
            <mix:PhotometricInterpretation>
                <mix:ColorSpace/>
            </mix:PhotometricInterpretation>
        </mix:Format>
        <mix:File>
            <mix:ImageIdentifier>www.aniagolu.org/images/people_dallas.jpg</mix:ImageIdentifier>
            <mix:FileSize>189010</mix:FileSize>
        </mix:File>
        <mix:PreferredPresentation/>
    </mix:BasicImageParameters>
    <mix:ImageCreation/>
    <mix:ImagingPerformanceAssessment>
        <mix:SpatialMetrics>
            <mix:ImageWidth>1020</mix:ImageWidth>
            <mix:ImageLength>767</mix:ImageLength>
        </mix:SpatialMetrics>
        <mix:Energetics>
            <mix:BitsPerSample>8,8,8</mix:BitsPerSample>
        </mix:Energetics>
    </mix:ImagingPerformanceAssessment>
    <mix:ChangeHistory/>
</mix:mix>
```

METS Website Viewer



Why METS Profiles?

- As a transfer, archiving or functional syntax METS is a paradigm of flexibility.
- But METS' very flexibility is at once its strength and its onus.
- How can we constrain METS instance production to facilitate interoperability?
- Enter METS Profiles

Profiles as Blueprints for Classes of Objects

- METS XML Schema and Profile Schema expressly written in WXS to take advantage of O-O modeling
- Profiles, by specifying constraints beyond those of the METS schema proper, create classes of objects whose properties are predictable

XML Schema Constraints

- Structural validation of the document and its elements' content model
  - does it have a structMap?
- Referential and identity integrity
  - Do the IDs and IDREFs match up? Are IDs unique?
- Data type validity
  - Is the string value kb or b **not** appended to your mix:fileSize?

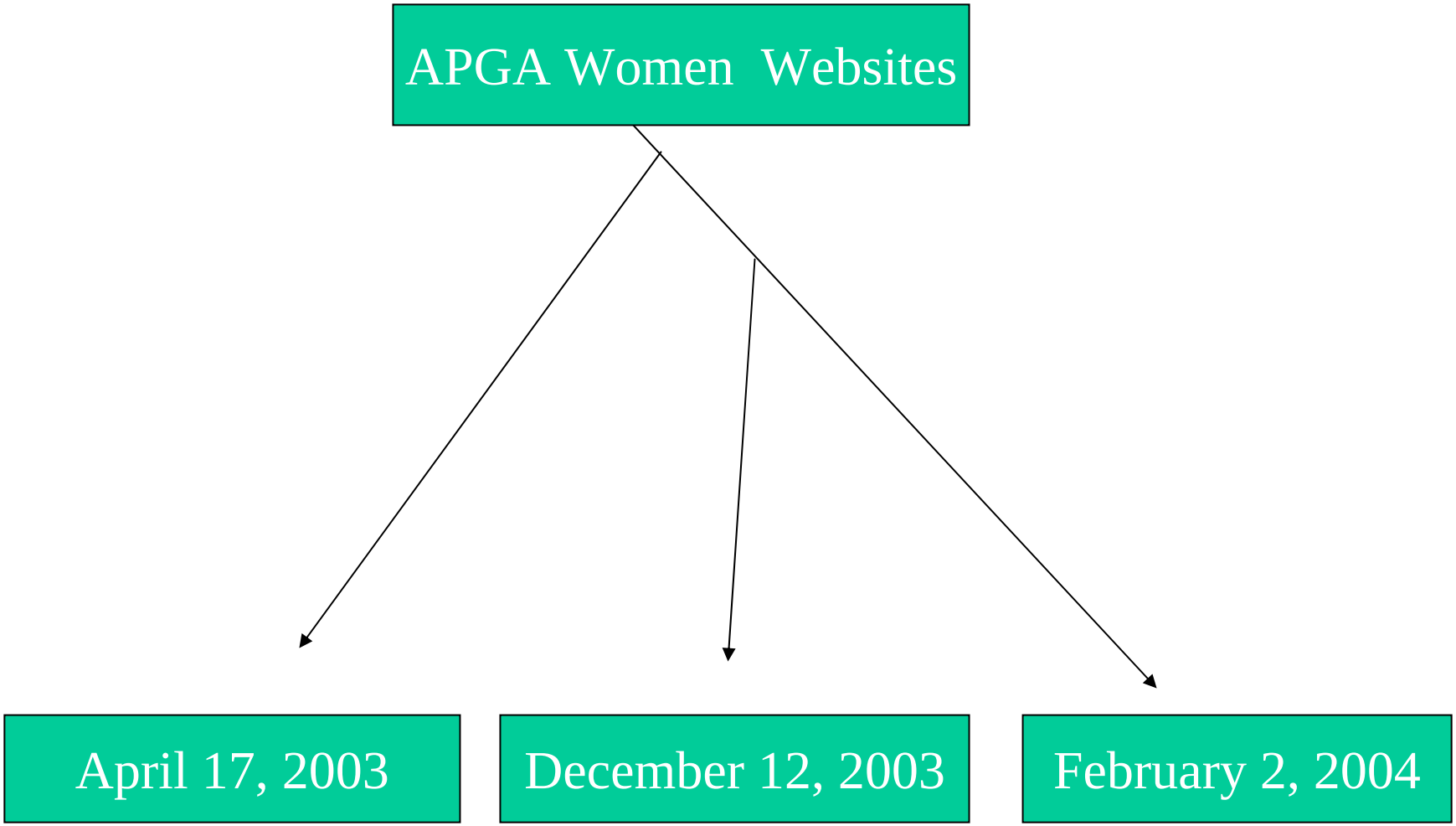
Profile Constraints Go Further

- May specify particular extension schema, rules of description, or controlled vocabs
- May specify arrangement and use of METS elements and attributes
- May specify technical characteristics of data files within a METS object
- May prescribe particular tools or applications to be used with a METS object

Modeling Web Site Object(s) in a continuous archive

One possibility for PWCA:

- Root level node (web site in the abstract) with <mptr>s to
- Intermediary node (web site harvested on April 17, 2003) with <mptr>s to
- Leaf node (single web page in web site harvested on April 17, 2003)


APGA Women Websites
April 17, 2003
December 12, 2003
February 2, 2004

LC Profile / Model

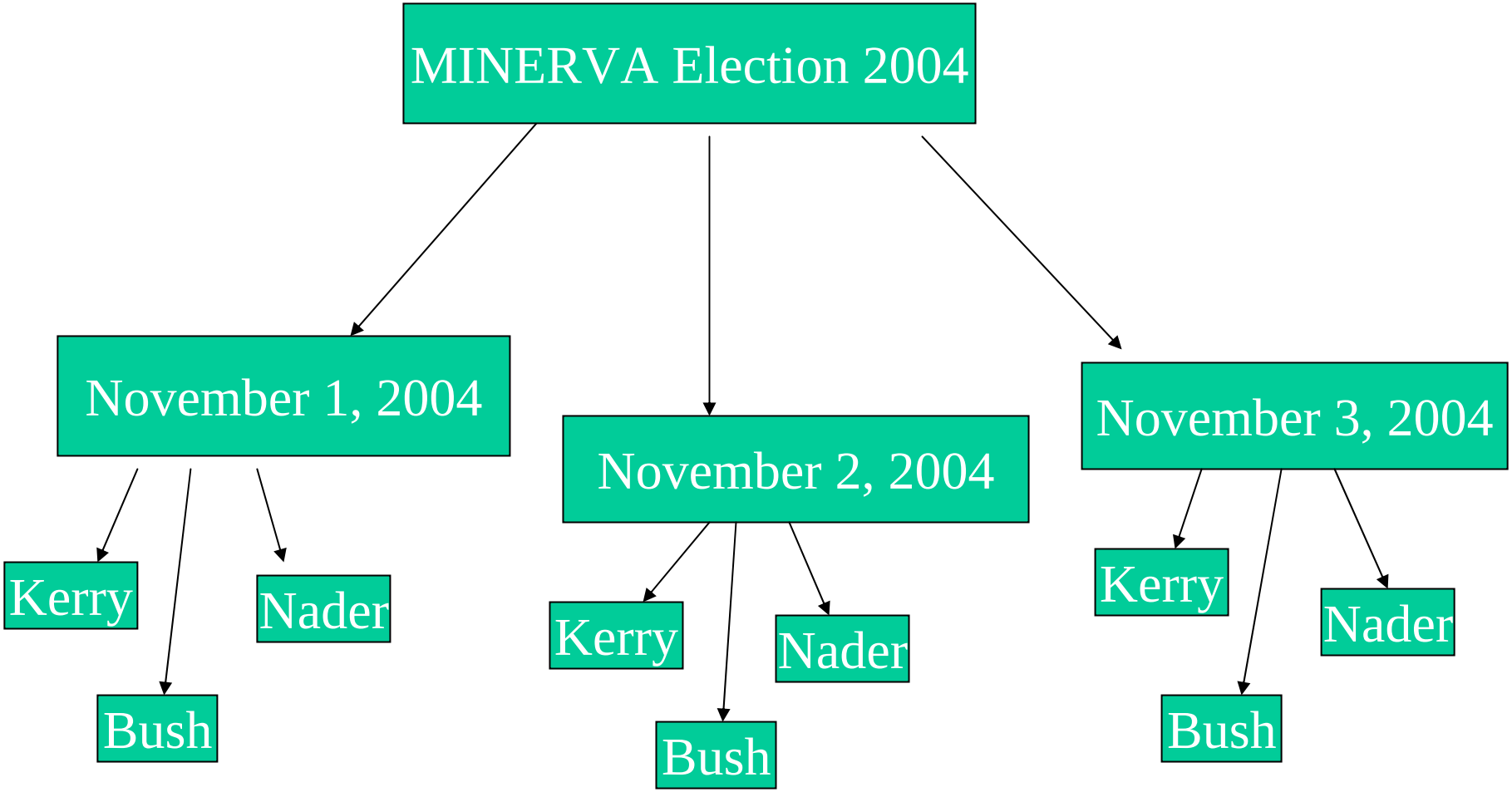
- Describing Internet Libraries e.g. MINERVA with different configurations of aggregation of content
- Profile and model to serve as “superclass” for other web site profiles e.g. PCWA

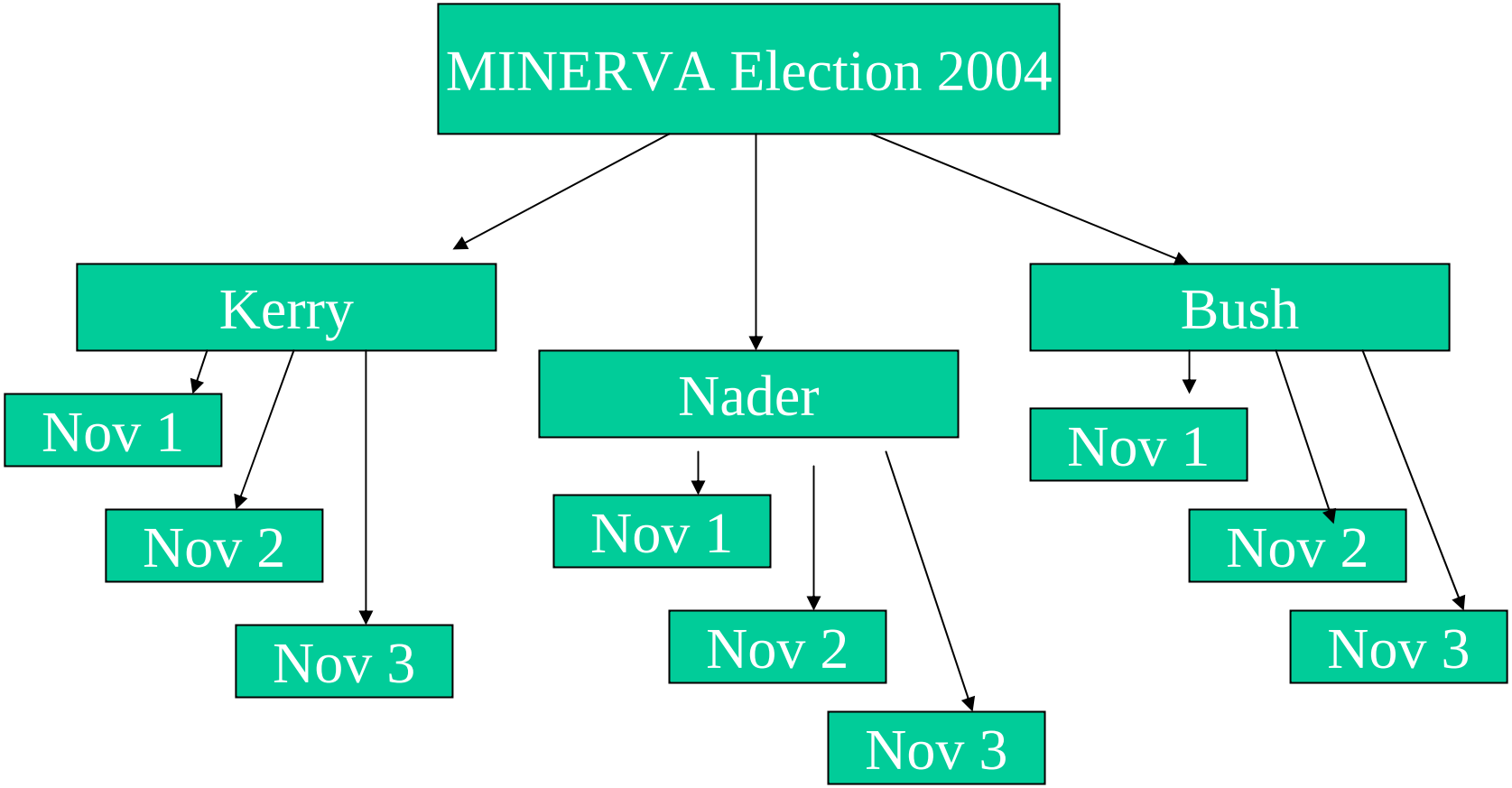
Aggregator and Single Captures Model

- Profile for top level aggregation that uses <mptr>s to point to either another intermediary aggregator or to more than one captured version of a web site.
- Profile for single standalone captured site, whether part of successive harvests or a one-off capture.

Aggregator Profile

- Contains single MODS description describing the aggregation as an intellectual object
  - e.g. Election 2004; JohnKerry.com (Nov 1-10)
- Contains no amdSec, fileSec or structLink.
- Contains a root for the aggregation nesting s with <mptr>s to each subsidiary aggregation or captured version.


MINERVA Election 2004
November 1, 2004
November 2, 2004
November 3, 2004
Kerry
Bush
Nader
Kerry
Bush
Nader
Kerry
Bush
Nader


MINERVA Election 2004
Kerry
Nov 1
Nov 2
Nov 3
Nader
Nov 1
Nov 2
Nov 3
Bush
Nov 1
Nov 2
Nov 3

Single Capture Profile

- Top-level MODS description describes the web site as it existed when it was captured.
- MODS record for each HTML page.
- amdSec uses textMD, MIX, A/V prototype
- fileSec
- structMap
- structLink

Modeling web site structure(s)

- Flattened logical tree hierarchy in structMap
  - entry page as root nesting all other pages (LC)
  - optionally further nesting pages' hyperlinked pages (PWCA)
- is cross-referenced to hyperlink structure in structLink <smLink> element
  - either from page to page (LC) or
  - from link to page (PWCA)

http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/apgawomen20030417.xml - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/apgawomen20030417.xml Go Links

Back to Cross-Collection Search

Title: APGA Women
Date Captured: 20030417
Abstract: Supports the All Progressive Grand Alliance political party (APGA). Information on the APGA presidential candidate, Chief Chukwuemeka Odumegwu-Ojukwu. Based in Kennesaw, Georgia.

Links

[Back to Homepage]

home

Search for

Search Clear

CRL Political Web Archiving Project METS Viewer

http://dlibdev.nyu.edu/webarchive/metstest/www.apgawomen.org/

Open in New Window Open in Full Frame

Internet

start Eudora - [In] Microsoft PowerPoint ... euterpe.bobst.nyu.e... emacs@MARTHA http://dlibdev.nyu.ed... 10:14 AM

LC structure

```
<METS:div DMDID="DM01" TYPE="wc:website" ID="page18"
LABEL="http://dlibdev.nyu.edu/webarchive/metstest/www.apgawomen.org/">
    <METS:fptr>
        <METS:par>
            <METS:area FILEID="FID18"/>
            <METS:area FILEID="FID1036"/>
            <METS:area FILEID="FID1043"/>
            <METS:area FILEID="FID1075"/>
        </METS:par>
    </METS:fptr>


<METS:structLink>
        <METS:smLink from="page18" to="page1059"/>
        …
        <METS:smLink from="page1059" to="page154"/>
```

```
<METS:structMap>
   <METS:div DMDID="DM01" TYPE="web site" ID="page18"
                LABEL="http://dlibdev.nyu.edu/webarchive/metstest/www.apgawomen.org/"
ORDER="01">
                    <METS:fptr>
                                  <METS:par>
                  <METS:area FILEID="FID18"/>
                  <METS:area FILEID="FID1036"/>
                  <METS:area FILEID="FID1043"/>
                  <METS:area FILEID="FID1075"/>
              </METS:par>
                     </METS:fptr>
         <METS:div ID="LINK1" LABEL="home" ORDER="2">
                              <METS:fptr>
              <METS:area BEGIN="000" BETYPE="BYTE" END="111" FILEID="FID18"/>
            </METS:fptr>
         </METS:div>

    <METS:structLink>
            <METS:smLink from="LINK1" to="page1059"/>
            …
            <METS:smLink from="LINK2" to="page113"/>
            <METS:smLink from="LINK3" to="page120"/>
```

Sample Archive Record - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://dlibdev.nyu.edu:8083/xmldev/servlet/SaxonServlet?source=apgawomen-root.xml&style=modspage3.xsl Go Links

Sample Archive Record

Title:	Web Site of APGA Women
Abstract:	Supports the All Progressive Grand Alliance political party (APGA). Information on the APGA presidential candidate, Chief Chukwuemeka Odumegwu-Ojukwu. Includes calendar, officers, membership information, projects page, news archive. Based in Kennesaw, Georgia.
Capture Date Range:	20030417-20040202
Subjects:	Political Parties
Language:	eng
Genre:	Web site
Access Condition:	[Policy Statement Goes Here]
Active URL:	http://www.apgawomen.org
Archive:	CRL Political Web Archiving Project

Go to METS Viewer for Archived Versions of This Site:

April 17, 2003 December 12, 2003 February 2, 2004

Internet

start Eudora - [In] Microsoft PowerPoint ... euterpe.bobst.nyu.e... emacs@MARTHA Sample Archive Recor... 10:34 AM

http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/apgawomen20030417.xml - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://dlibdev.nyu.edu:8083/xmldev/servlet/frames/apgawomen20030417.xml Go Links

Back to Cross-Collection Search

Title: APGA Women
Date Captured: 20030417
Abstract: Supports the All Progressive Grand Alliance political party (APGA). Information on the APGA presidential candidate, Chief Chukwuemeka Odumegwu-Ojukwu. Based in Kennesaw, Georgia.

Links

[Back to Homepage]

home

Search for

Search Clear

CRL Political Web Archiving Project METS Viewer

http://dlibdev.nyu.edu/webarchive/metstest/www.apgawomen.org/

Open in New Window Open in Full Frame

Internet

start Eudora - [In] Microsoft PowerPoint ... euterpe.bobst.nyu.e... emacs@MARTHA http://dlibdev.nyu.ed... 10:14 AM