PORTICO

Long-Term Preservation and Standards: An Uneasy Alliance

Evan Owens
Chief Technology Officer
evan.owens@portico.org
www.portico.org

This Presentation

- Prologue
- Part I: Digital Preservation and Standards
  - Lighting Overview of Preservation, Standards
  - Key Preservation Standards and Concepts
  - Limitations of Standards
- Part II: Some Standards War Stories
  - MS Office 2003 SP 3 & Obsolete Formats
  - TIFF & PDF/A
  - MS IE 8.0 & Web Standards
- Part III: Portico's Experiences in Digital Preservation
  - Portico Business & Technology Summary
  - Current State of E-Journal Publishing Practices & Standards
- Epilogue
  - Some Musicological Metaphors
  - Visions from the Past and Questions for the Future

PROLOGUE

Long term preservation requires that we store content on at least three continents, using three different operating systems, and under three different political systems

— Dale Flecker, Harvard University Library

PART I

Digital preservation and Standards

Preservation of Digital Objects

- Ensuring Long-Term Viability (Usability)
- 20, 50, 100 Years From Now, Can We
  - read the files?
  - understand the structure of the files?
  - be sure that we have an authentic copy of the work?
- Layers
  - Physical Layer: storage media
    - *Necessary but not sufficient*
  - Logical Layer: file formats, structured data
    - *Significant properties*
  - Conceptual/Intellectual Layer: the “work”
    - *Multiple representations over time*
- Approaches to Preservation
  - Emulate (or maintain) the original technology
  - Migrate (and/or normalize) to currently supported formats
  - Byte preserve for future digital archeologists
- “Interoperability with the Future”
  - Can we interoperate today?

Digital Preservation is Everyone's Problem

- Cultural Memory Institutions
  - Libraries, Archives, Museums
- Research Organizations
  - Universities, Laboratories, Data Centers
- Government Agencies
  - US GPO, NARA
- Corporations
  - Regulatory compliance, Business continuity
- Private Individuals
  - Personal digital content
  - Collections of licensed and free content

But it isn't exactly the same problem from everyone!

Varieties of Digital Preservation Projects

- Library and other digitization projects
  - Controlled environments; potential for good metadata
- Web site harvesting
  - Uncontrolled environment; minimal metadata available
- Electronic records and business data
  - Potential for lots of control; mandatory metadata and formats
- Published electronic content
  - Semi-controlled; good descriptive metadata; variable or no technical metadata
- Scientific data
  - Enormous quantities
  - High expectations for long-term usability

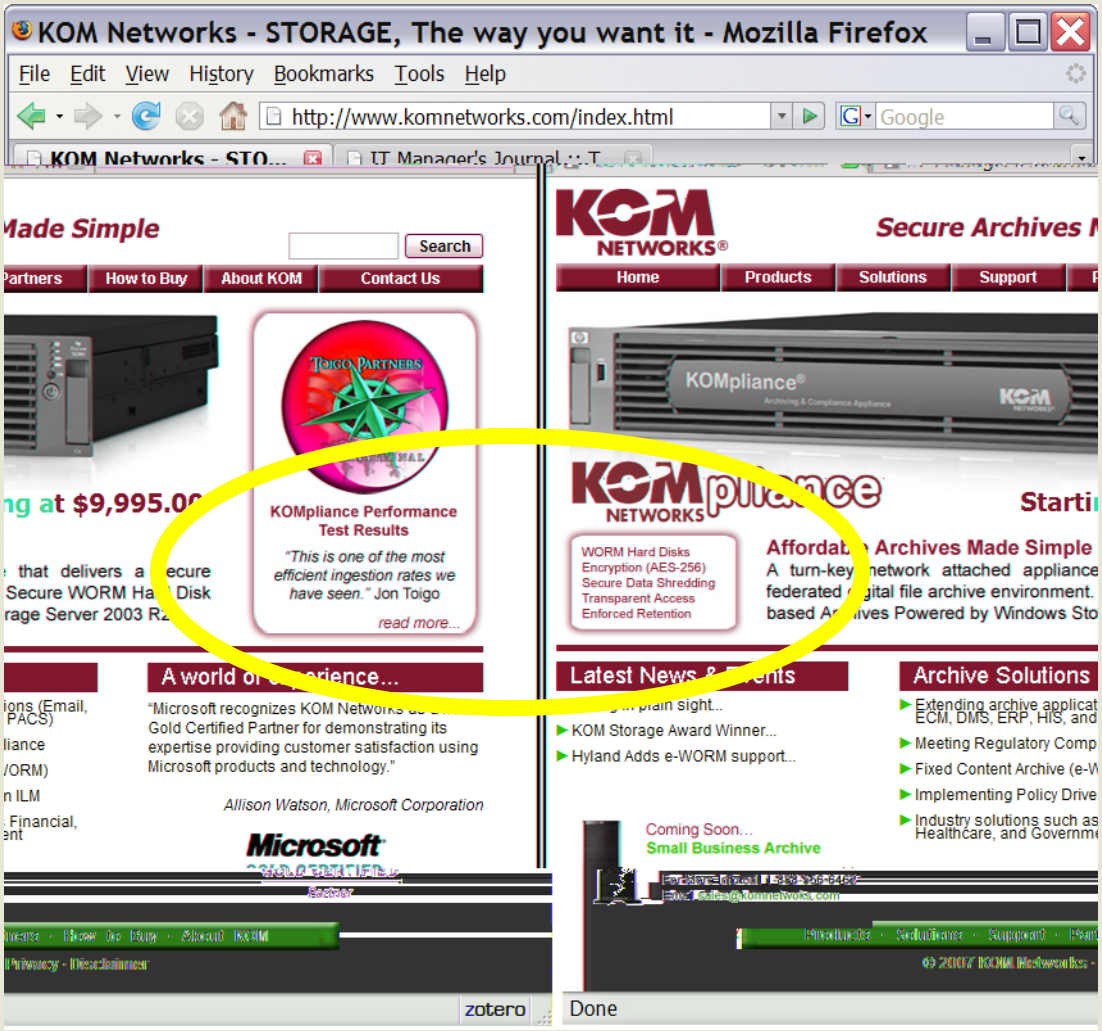
One size solution will not fit all!

Explosion of Digital Preservation Activities

- Traditional Players
  - National archives and libraries
  - Digital library community (e.g., CNI, DLF, JCDL, ECDL)
- New Organizations and Projects
  - Digital Preservation Coalition (UK, 2001)
  - Erpanet (EU, 2002?); now DigitalPreservationEurope
  - Digital Curation Center (UK, 2004)
  - National Digital Information Infrastructure and Preservation Program (US)
- Courses, Workshops, Conferences
  - Erpanet, DCC, DPC, Cornell, etc.
  - IS&T Archiving Conferences (from 2004)
  - International Conference on Preservation of Digital Objects (from 2004)
  - International Digital Curation Conference (from 2005)

This is all pretty recent!

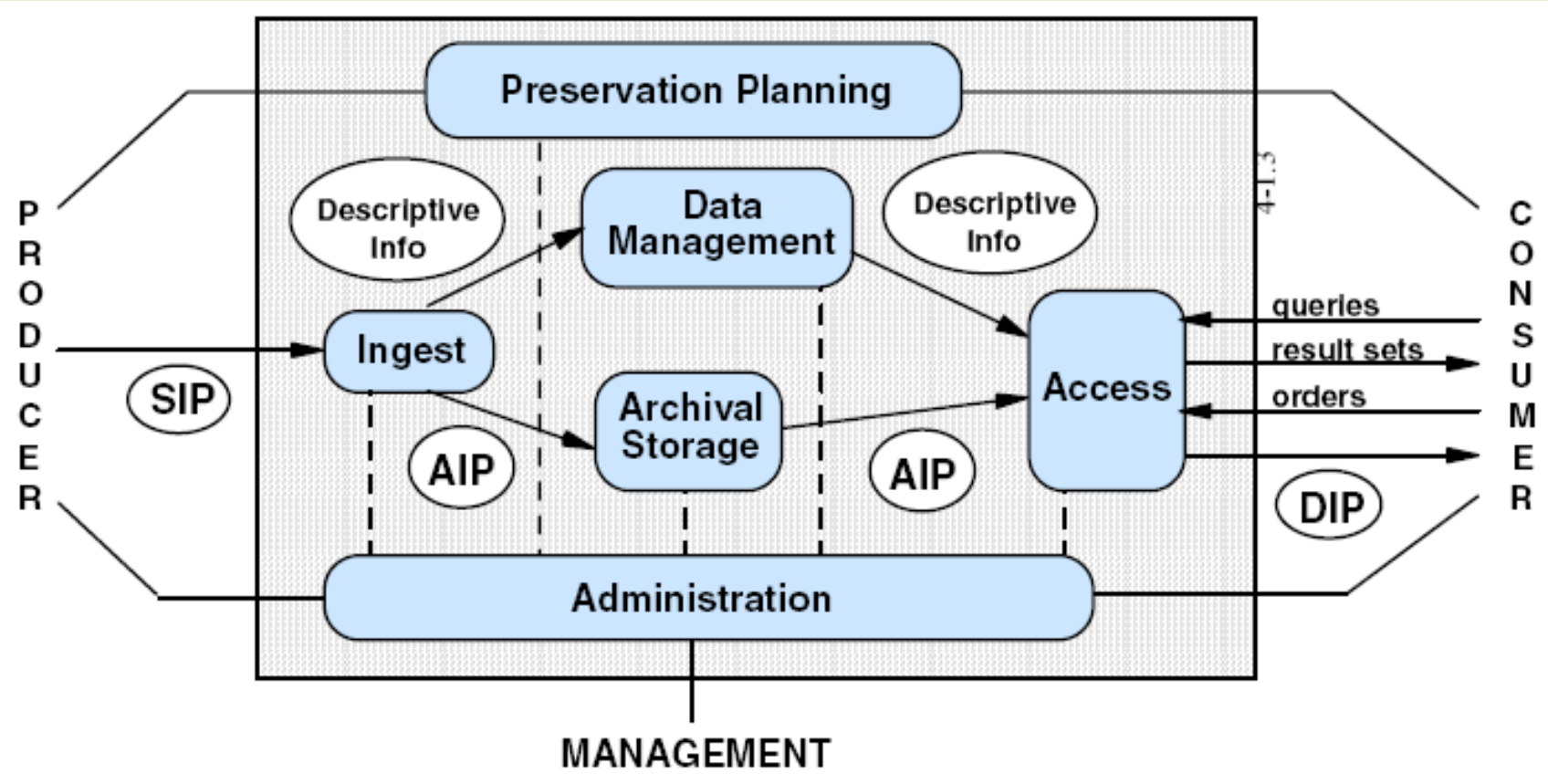
One of Many Commercial Solutions…But to Which Problem?



Some Milestones in the Short History of Digital Preservation

- Garrett, John, and Don Waters. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Task Force on Archiving of Digital Information, 1996. 64. http://www.rlg.org/ArchTF/
- *Reference Model for an Open Archival Information System - OAIS.* National Aeronautics and Space Administration - Consultative Committee for Space Data Systems, 2002. 148. http://public.ccsds.org/publications/archive/650x0b1.pdf
- *Trusted Digital Repositories: Attributes and Responsibilities.* Mountain View, CA:RLG-OCLC, 2002. 70. http://www.rlg.org/longterm/repositories.pdf
- *An Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment*. Mountain View, CA:RLG and NARA, 2005. 65. http://www.rlg.org/en/page.php?Page_ID=20769

The Ubiquitous OAIS Functional Model Diagram



Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1 (BLUE BOOK), January 2002, page 4-1.

OAIS Goes Commercial: from a Hitachi Brochure

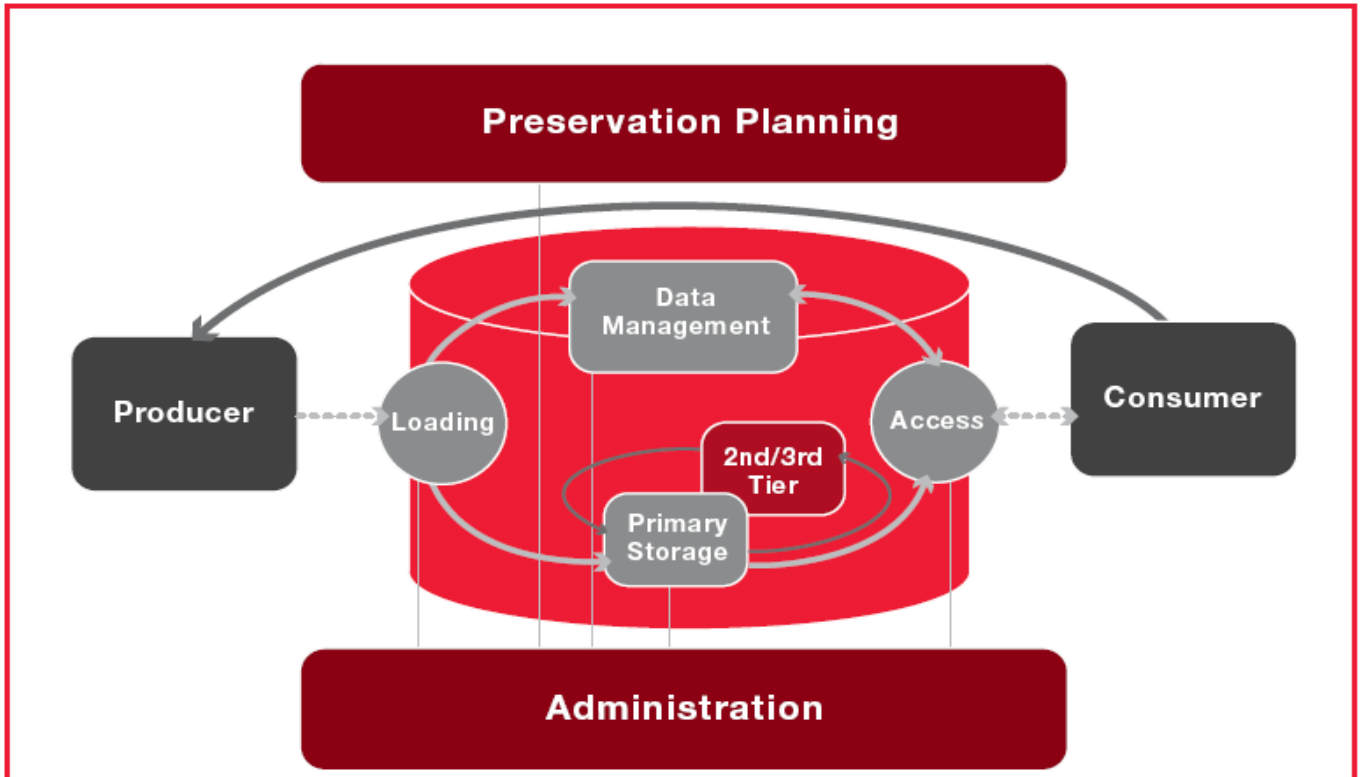
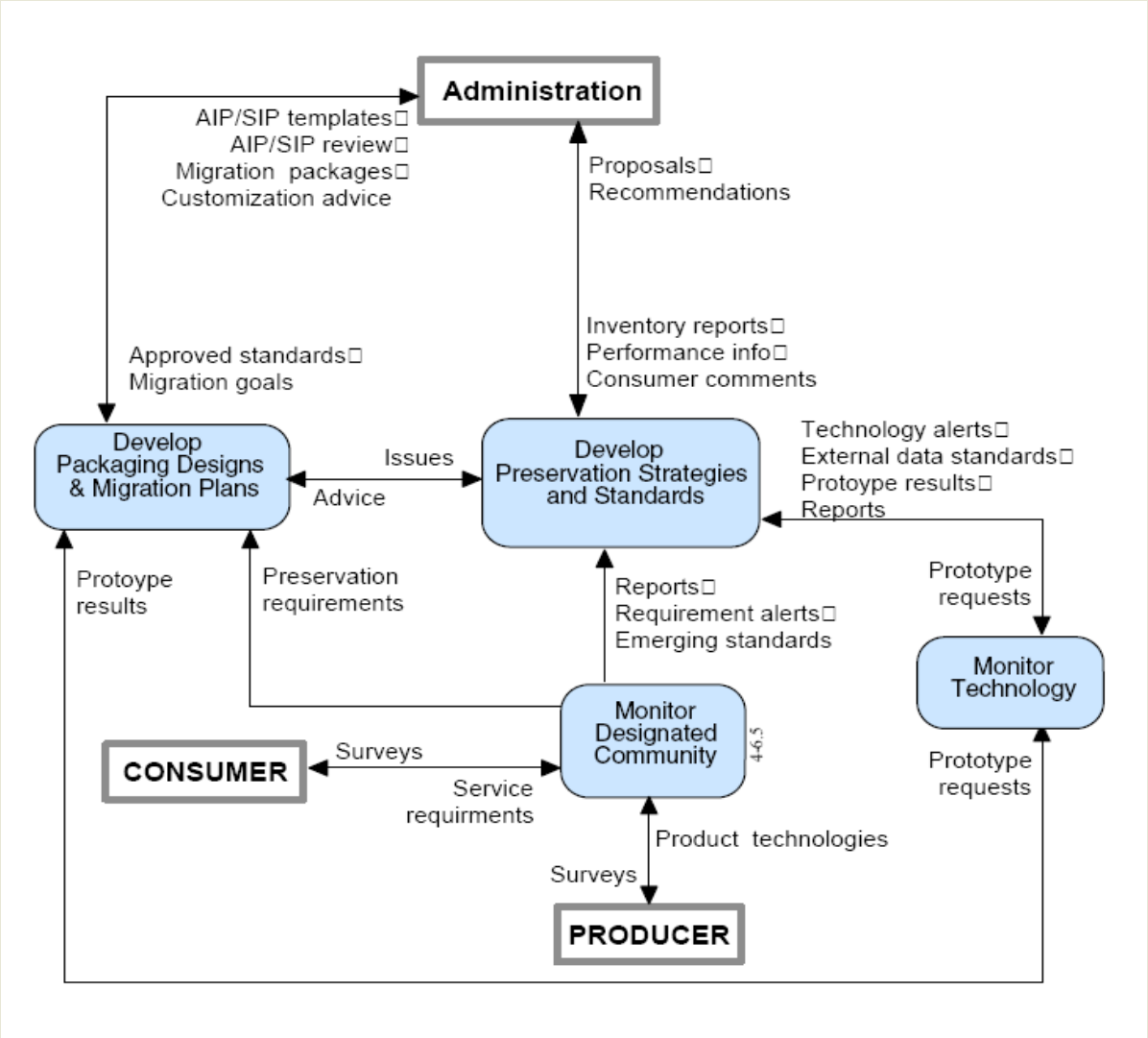
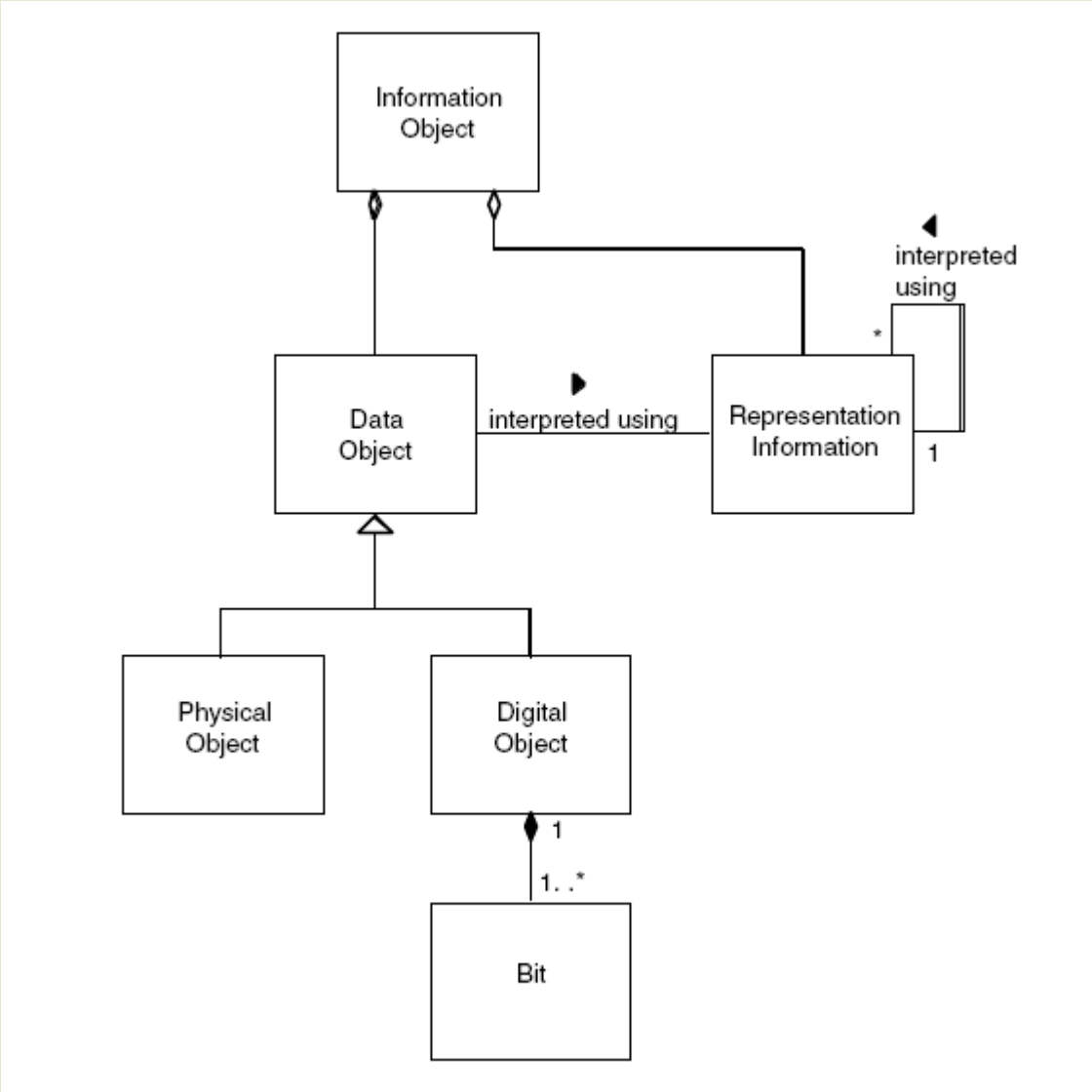


Figure 2. The Hitachi Data Systems Common Model for "Active Archive" solutions leverages the ISO Reference Model for Open Archival Information Systems with modifications to describe an active content flow.

OAIS Preservation Planning



OAIS Information Model



Select Technical Features of Preservation

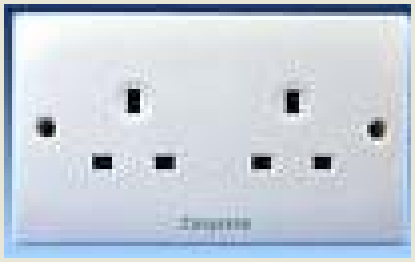
From *Preservation in the Age of Large-Scale Digitization*, Oya Y. Rieger, CLIR, February 2008

- Creation of a repository model to ingest, monitor, manage, and archive digital objects and associated metadata, files, and scripts
- Development and implementation of an ingest workflow and quality control measures to verify authenticity and completeness of ingested content
- Creation and management of preservation metadata (including technical metadata)
- Identification of properties to preserve in digital objects
- Continuous monitoring and management of digital content to detect bit corruption, loss, or obsolescence
- Record of provenance and change history for all objects

Select Technical Features of Preservation, Cont.

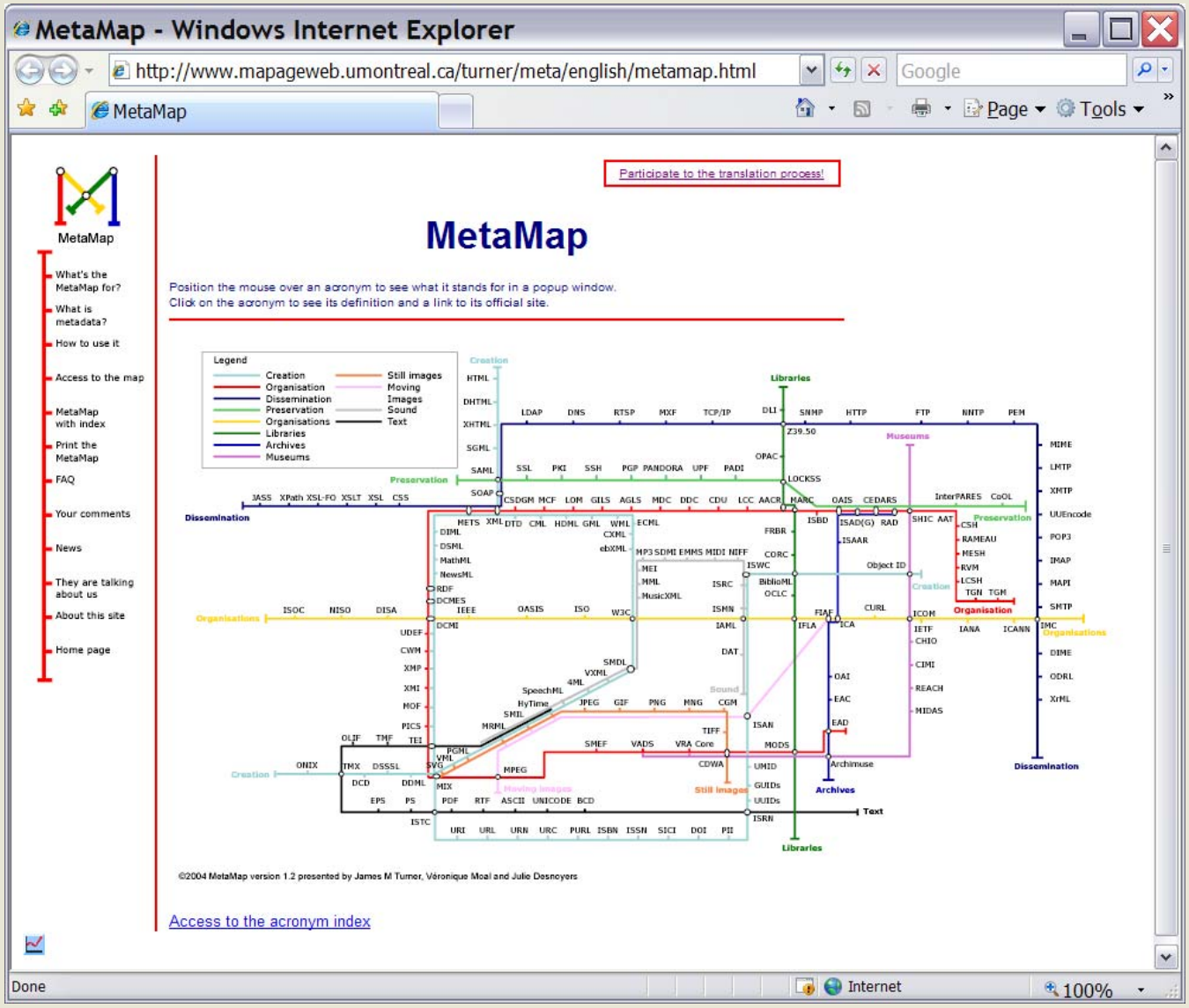
- Programs in support of various preservation strategies, including refreshing, migration, replication, normalization, and emulation (both for preventive measures and for staying abreast of standards and technologies)
- Disaster-prevention, recovery, and contingency plans
- Periodic review and updating of preservation procedures
- Mechanisms for monitoring triggers for preservation action (e.g., file format migration, file corruption)
- Security measures
- Technical audits

Standards are Great: Everyone Should Have One!



A Taxonomy of Standards

- Political Context of Standards
  - Voluntary (de facto) or Mandatory (de jure)
  - Public, Proprietary, Open
- Types of Standards (from Wikipedia article on standards)
  - Specification for Item, Material, Component, System, or Service
    - *File formats, Fixity checks, Metadata*
  - Test Methods
    - *Format validation*
  - Procedure or Practice
    - *Trusted Repositories checklist*
  - Definition or Terminology
    - *OAIS Reference Model*
- Role of Standards
  - Codify existing best practices
    - *Acid-free paper*
  - Enable new practices or technologies
    - *Web standards*


MetaMap - Windows Internet Explorer
http://www.mapageweb.umontreal.ca/turner/meta/english/metamap.html
MetaMap
Participate to the translation process!
MetaMap
Position the mouse over an acronym to see what it stands for in a popup window.
Click on the acronym to see its definition and a link to its official site.
What's the MetaMap for?
What is metadata?
How to use it
Access to the map
MetaMap with index
Print the MetaMap
FAQ
Your comments
News
They are talking about us
About this site
Home page
Legend
Creation
Organisation
Dissemination
Preservation
Organisations
Libraries
Archives
Museums
Still images
Moving Images
Sound
Text
©2004 MetaMap version 1.2 presented by James M Turner, Véronique Moal and Julie Desnoyers
Access to the acronym index
Done
Internet
100%


MetaMap - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.mapageweb.umontreal.ca/turner/meta/english/metan
Participate to the translation process!
MetaMap
Position the mouse over an acronym to see what it stands for in a popup window.
Click on the acronym to see its definition and a link to its official site.
XML Parsing Error: prefix not bound to a namespace
Location: http://www.mapageweb.umontreal.ca/turner/meta/english/me
Line Number 309, Column 1:
<a id="jass" xlink:href="dissemination.html#jass" target="blank">
^
Done
zotero

File Format Standards

- Standard Names for Formats?
  - Promise of the Global Digital Format Registry project
- Format Contexts
  - Mass Market: General Public
    - *Documents (MS Office, PDF), Images, Media*
  - Broad: Cross Industry
    - *Geo-spatial data, CAD*
  - Narrow: Industry-Specific
    - *Industry DTDs & Schemas*
    - *Discipline-Specific Scientific data formats*
  - Very Narrow: Company-Specific
    - *Proprietary DTDs & Schemas*
- An Ideal Format Standard would
  - have two independent implementations
    - Create objects according to the standard
  - have two independent validations
    - Verify conformance of objects to the standard
  - be freely available
  - be less than 1000 printed pages

Limitations of Standards in Preservation

- Technology standards don't stand still
  - SGML in 1986
  - XML 1.0 in 1998
  - XML 2.0?
- Technology standards are often driven by commercial motivations
  - Change for change's sake
    - Feature bloat
  - Planned obsolescence
  - Battle of the market place
    - MS Word versus WordPerfect
- Standards creation can be immensely political
  - E.g., OpenXML, IE 8 and web standards
- Measuring standards conformance isn't always easy
  - Validity doesn't always equate to usability
- Standards and applications don't always agree
  - HTML browsers, Acrobat Reader

A Personal Reminder of the Evolution of Standards: My Computing Environment circa 1985

- Computer: Epson QX-10
- CPU: Z80 4 MHz 256K RAM
- Operating System: CP/M-80
- Storage: 5-1/4” floppy disk
- Output: continuous-feed paper, daisy-wheel printer
- Software: Wordstar 3.x, dBASE II

http://www.obsoletecomputermuseum.org/qx-10/

PART II

Some Standards War Stories

Standards War Stories: MS Office 2003 SP 3

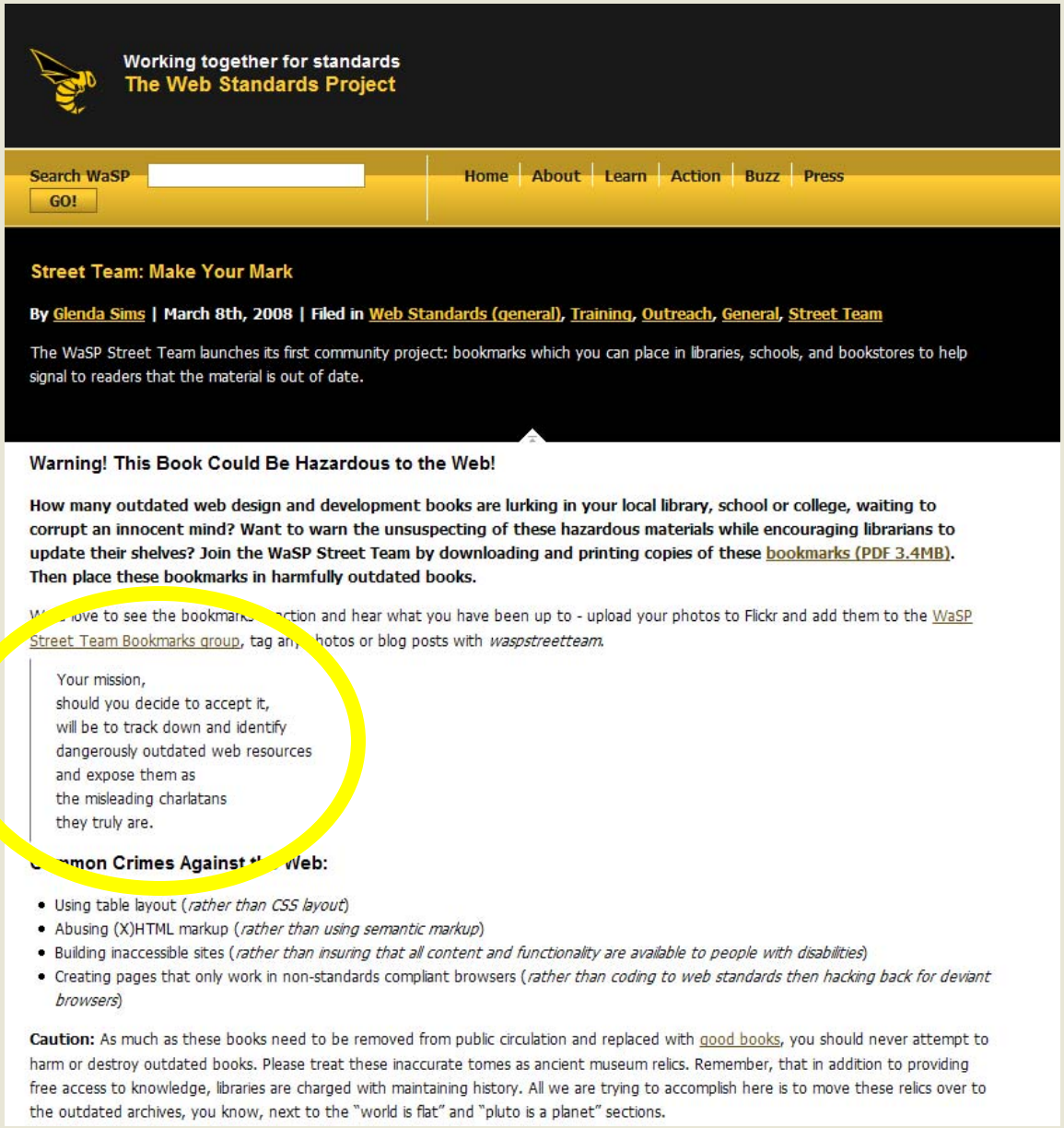
- MS Office 2003 Service Pack 3 prevents opening older formats
- Blocked formats:
  - MS Word before 6.0
  - Lotus and Quattro spreadsheet formats
  - Corel Draw
  - Excel 4.0 Charts
  - dBASE II
  - PowerPoint prior to PowerPoint 97
- Full details at http://support.microsoft.com/kb/938810
  - A patch to undo the change is available
- Is this the beginning of the end of MS Office support for older formats?
- Library of Congress has asked Microsoft to comment
- OAIS Preservation Planning model suggest that we should begin thinking about migrations for these formats

Standards War Stories: TIFF and PDF/A

- Portico executing a planned migration test of TIFF to PDF/A
  - Time frame was Summer 2007; all results now obsolete
  - Research question:
    - What conversion tool produces valid PDF/A?
    - What validation tool can correctly identify the results as PDF/A?
- Conversion tools
  - Four tools tested, none created valid PDF/A
  - One vendor agreed to fix problem; released new version
- Validation tools
  - Four tools tested, only one correctly identified the problems
- Key problem
  - Design weakness in PDF that allows key metadata to be stored twice
  - Ambiguity in the PDF/A specification about how closely the two sets of metadata must match
- Surprise result
  - Adobe Acrobat 8 behaves badly: it rewrites what is in the file without warning

Standards War Stories: MS IE 8 & Web Standards

- Microsoft Internet Explorer is famous for standards non-compliance
- Should IE 8.0 default to standards or to IE 7.0's non-standard behavior?
  - Enormous fuss in community
  - Microsoft changed its position
- From their press release:
  - Consistent with its efforts to promote further interoperability across the Web, Microsoft Corp. is now configuring the settings in Internet Explorer 8, the upcoming version of its browser, to render content — by default — using methods that give top priority to Web standards interoperability. …
  - "This is obviously a complex issue, with important considerations on both sides," Ozzie said. "On one hand, there are literally billions of Web pages designed to render on previous browser versions, including many sites that are no longer actively managed. On the other hand, there is a concrete benefit to Web designers if all vendors give priority to interoperability around commonly accepted standards as they evolve. After weighing these very legitimate concerns, we have decided to give top priority to support for these new Web standards.



Working together for standards
The Web Standards Project

Search WaSP GO!

Home | About | Learn | Action | Buzz | Press

Street Team: Make Your Mark

By Glenda Sims | March 8th, 2008 | Filed in Web Standards (general), Training, Outreach, General, Street Team

The WaSP Street Team launches its first community project: bookmarks which you can place in libraries, schools, and bookstores to help signal to readers that the material is out of date.

Warning! This Book Could Be Hazardous to the Web!

How many outdated web design and development books are lurking in your local library, school or college, waiting to corrupt an innocent mind? Want to warn the unsuspecting of these hazardous materials while encouraging librarians to update their shelves? Join the WaSP Street Team by downloading and printing copies of these bookmarks (PDF 3.4MB). Then place these bookmarks in harmfully outdated books.

W[illegible]ove to see the bookmark[illegible]ction and hear what you have been up to - upload your photos to Flickr and add them to the WaSP Street Team Bookmarks group, tag an[illegible]hotos or blog posts with *waspstreetteam*.

> Your mission,
> should you decide to accept it,
> will be to track down and identify
> dangerously outdated web resources
> and expose them as
> the misleading charlatans
> they truly are.

C[illegible]mon Crimes Against t[illegible] Web:

- Using table layout (*rather than CSS layout*)
- Abusing (X)HTML markup (*rather than using semantic markup*)
- Building inaccessible sites (*rather than insuring that all content and functionality are available to people with disabilities*)
- Creating pages that only work in non-standards compliant browsers (*rather than coding to web standards then hacking back for deviant browsers*)

Caution: As much as these books need to be removed from public circulation and replaced with good books, you should never attempt to harm or destroy outdated books. Please treat these inaccurate tomes as ancient museum relics. Remember, that in addition to providing free access to knowledge, libraries are charged with maintaining history. All we are trying to accomplish here is to move these relics over to the outdated archives, you know, next to the "world is flat" and "pluto is a planet" sections.

http://www.webstandards.org/2008/03/08/street-team-make-your-mark/

PART III

Portico's Experiences in Standards-Based Digital Preservation

Portico Business Summary

- A permanent archive of scholarly literature in electronic form
  - All preservation and access rights secured by irrevocable contractual agreements
- Initial content area is E-Journals
  - 50 participating publishers
  - 7,334 journal titles committed
    - 5,562 titles with post cancellation access through Portico
  - 425 participating libraries from 11 countries
  - 6,142,211 articles archived; >14M articles committed
    - > 60 Million files
    - 94 file formats
- Currently ingesting an average of 1 million articles per month
  - 10 million files
  - 100 GB of METS/PREMIS/JHOVE metadata per month
- Start-up funding by Andrew W. Mellon Foundation, JSTOR, Ithaka, and Library of Congress NDIIPP

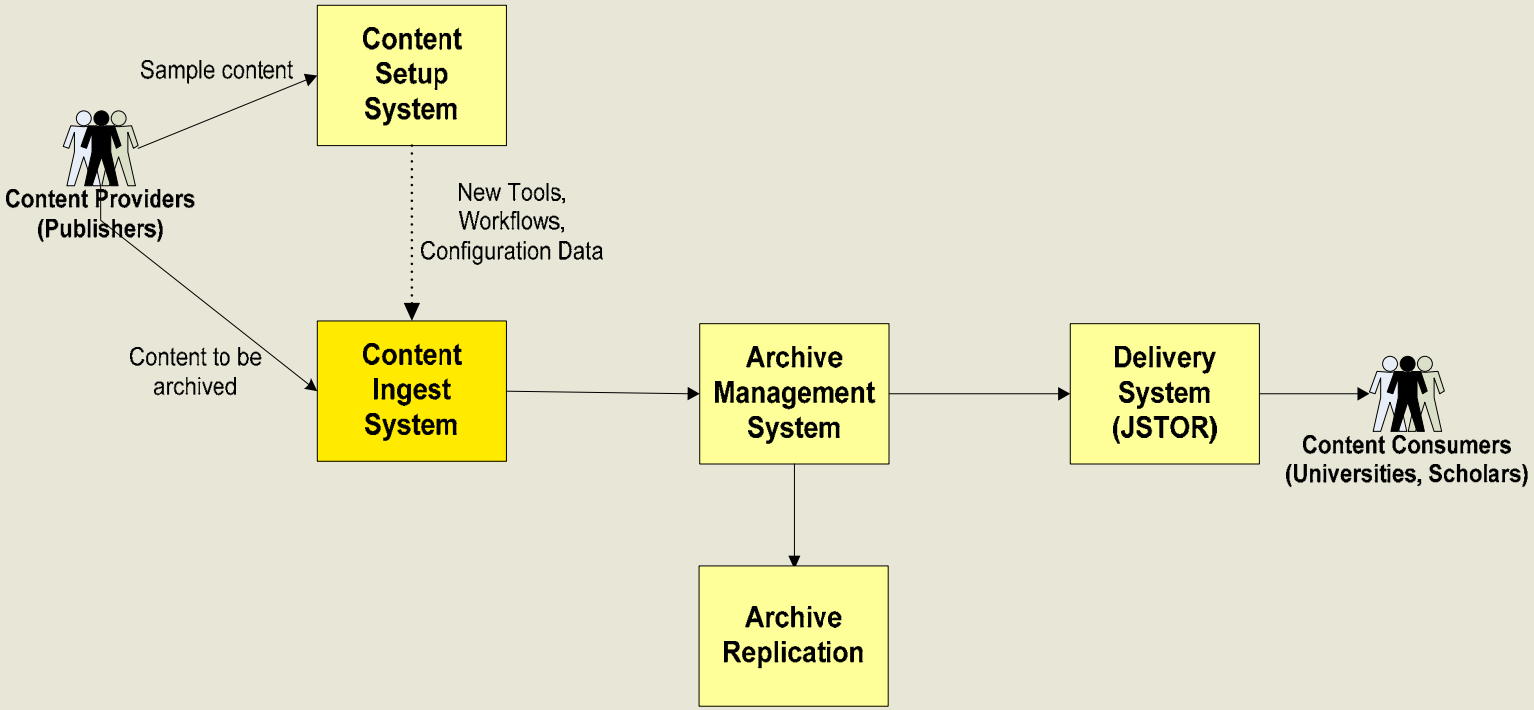
Portico Technology Summary

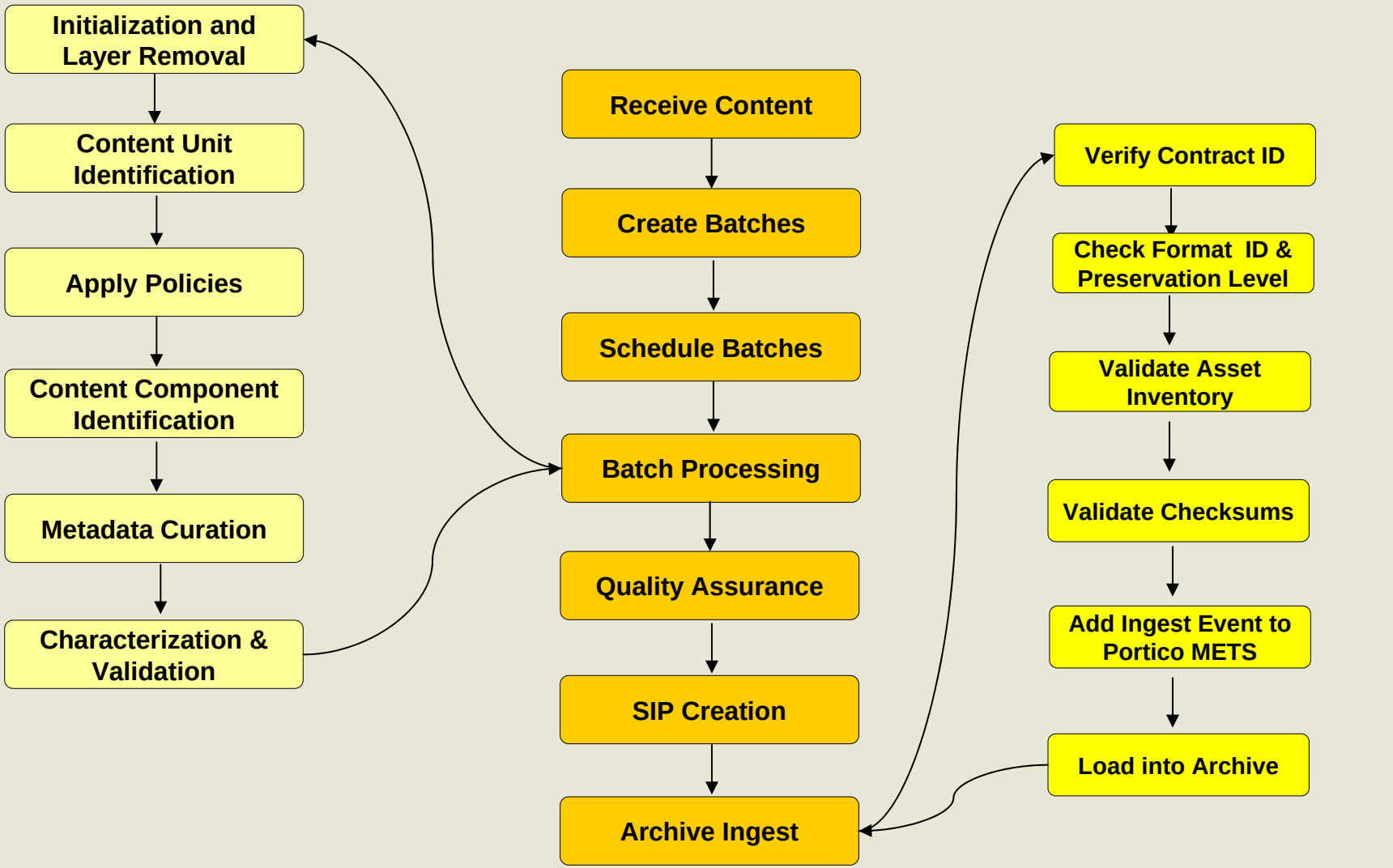
- OAIS-compliant repository designed for managed preservation
- Key influences:
  - OAIS, GDFR, PreMIS, METS, MPEG-21, ARK
- Key technologies:
  - XML, XML schema, Schematron, JHOVE, NOID
  - Documentum, Oracle, Java, JMS, LDAP
  - Format Registry
- Archive design goals:
  - Content preserved in application-neutral content using open standards
    - METS, PREMIS, JHOVE
  - A “Bootstrapable Archive”
    - XML plus digital objects
- Ingest system design goals:
  - Pluggable tools to facilitate new providers and replacement tools
  - Configurable workflows for different content types
  - Scalable to very high content volumes

Portico Preservation Policies

- Format-based migration strategy
  - Driven by Portico Format Registry
- Preservation policies:
  - Fully supported
  - Reasonable effort
  - Byte-preserve only
- Preservation policies based on
  - Format validity
  - File format action plans and archive capabilities
  - Business rules such as publisher preferences
- Archive must also preserve supporting information
  - Required files such as DTDs and entity files
  - Documentation
  - Contracts
  - Archive policy documents
  - Archival actions documents

Portico Systems Overview




Initialization and Layer Removal
Content Unit Identification
Apply Policies
Content Component Identification
Metadata Curation
Characterization & Validation
Receive Content
Create Batches
Schedule Batches
Batch Processing
Quality Assurance
SIP Creation
Archive Ingest
Verify Contract ID
Check Format ID & Preservation Level
Validate Asset Inventory
Validate Checksums
Add Ingest Event to Portico METS
Load into Archive

Result: 100s of Gigabytes of Preservation Metadata

AACCA_PMETS_1.xml*

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PorticoMETS CONTENTIDS="ark:/27927/tc517f6d" CONTENTSET="ISSN_00014826"
    CONTENTTYPE="E-Journal Content" CONTENTUNIT="Article" OBJID="ark:/27927/tc530jdc"
    PROFILE="PorticoMETS 1.2 based on METS 1.4"
    xsi:schemaLocation="http://www.portico.org/standards/Public/XML/schema/PorticoMETS/1.2/ http://www.portico.
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns="http://www.portico.org/standards/Public/XML/schema/PorticoMETS/1.2/">
    <metsHdr CREATEDATE="2008-01-18T13:09:18.109-05:00">
        <agent ROLE="CREATOR">
            <name>Portico Content Preparation System v1.1.0</name>
        </agent>
    </metsHdr>
    <structMapContent>
        <div AGREEMENT="AACCA Agreement, Version 1.0, Apr1 12, 2007"
            AGREEMENTIDS="ark:/27927/ps011jg" CONTENTIDS="ark:/27927/tc517f6d" LABEL="Article"
            TYPE="Content Unit">
            <mdGroup>
                <descMDcurated CONTENTIDS="ark:/27927/tc517n06"
                    CREATED="2008-01-18T13:24:34.624-05:00" VERSION="1.0">
                    <mdWrap>
                        <xmlData>
                            <ns1:PorticoArticleMetadata
                                xsi:schemaLocation="http://www.portico.org/standards/Public/XML/schema/PorticoArtic
                                sort-key=" ISSN_00014826  83      1         157    "
                                schema-version="1.1" metadata-type="curated"
                                display-label="ISSN_00014826 v83 n1 p157-184"
                                xmlns:ns1="http://www.portico.org/standards/Public/XML/schema/PorticoArticleMetada
                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
                                <ns1:article xml:lang="en"
                                  article-type="research-article">
```

E-Journal Content Processing

- Inputs
  - Per article: one text or metadata file, zero or more other files
  - Arbitrary (publisher-specific) collections of data
    - Proprietary file & directory naming conventions
    - Proprietary formats
  - Undocumented business rules hidden in the data
- Outputs
  - Content packaged in Portico METS
  - Metadata: technical, descriptive, events
  - Content restructured to Portico content model
    - Article component structure documented
  - Content normalized as per preservation plans
    - Proprietary publisher DTDs converted to NLM Archival DTD
    - PDF created from TIFF as needed

E-Journal Publishing Environment

- Publisher system changes
  - To new delivery vendors
  - To new delivery platforms
- Publisher content changes
  - To NLM DTD
  - Adding digitized print
- Variety of content formats per publisher
  - Current e-journals
  - Older e-journals
  - Digitized print
  - Averaging 2.4 per publisher thus far; will go down
- Consultative role for Portico
  - Publishers eager for outside comments on production issues
  - Exchanging data brings out questions & identifies weaknesses
    - Especially for publishers who do not currently exchange data

E-Journal Content: Good, Bad, and Ugly

- Portico's automated and manual quality control has revealed a wide range of problems with published e-journal content, most of which were also wrong on the publisher's web site
- Bad Files
  - TIFF, JPEG, GIF
  - PDF problems of all sorts
  - XML not parsed
  - XML tag abuse
- Missing Files
  - Graphics
- Missing Articles
  - Missing from publisher web sites
- Mangled Metadata
  - CrossRef has this problem also

E-Journal Content Management is relatively new; lots of room for improvements in practices and standards.

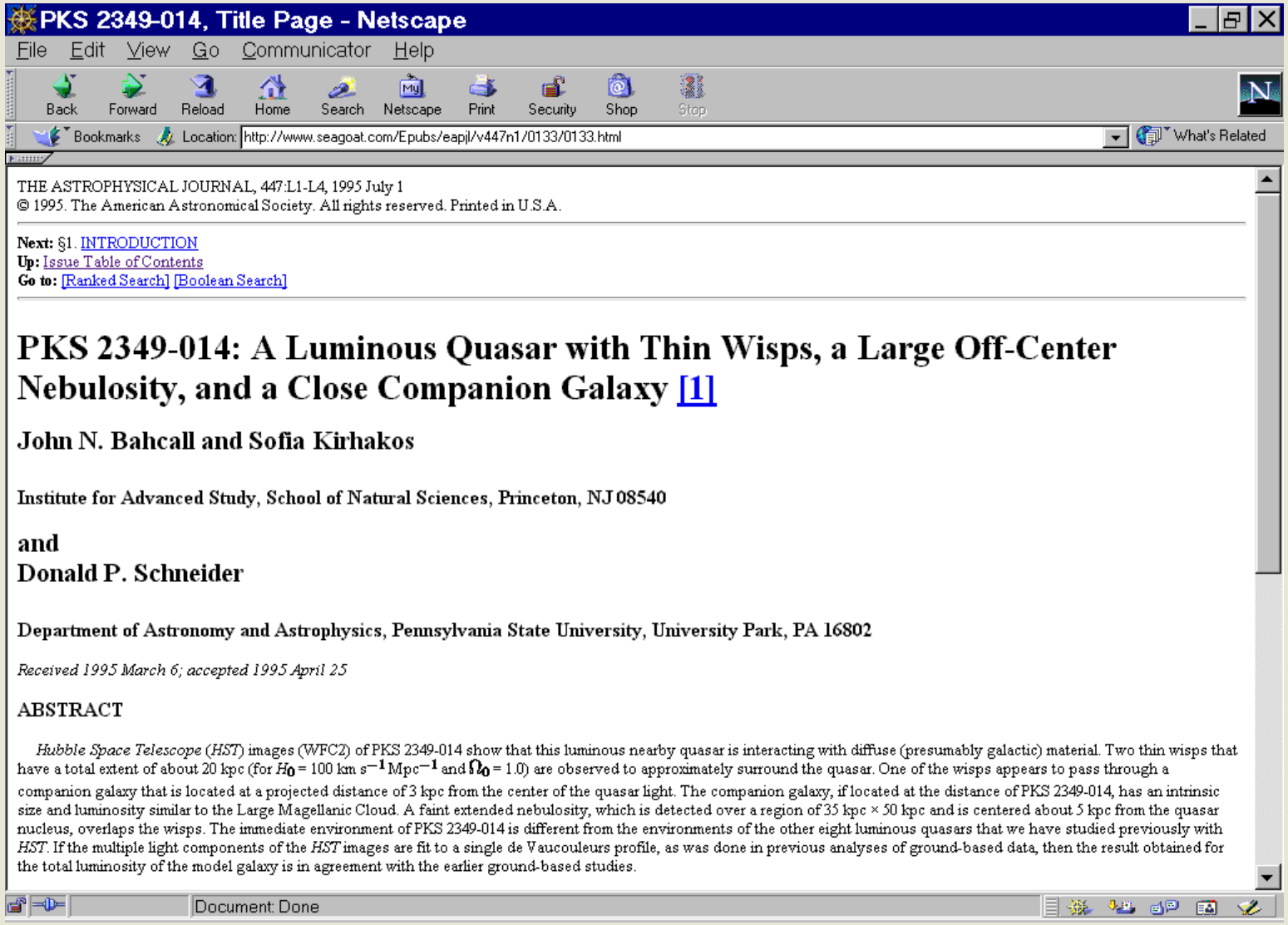
E-Journal Publishing Problem Areas: Standards or Lack Thereof

- Content management and quality control
  - Documentation, naming, packaging
    - *No real standard except perhaps Elsevier dataset.toc*
  - Production content: PDF, XML, graphics
    - *NLM Tag set for XML and emerging standard*
  - Author-supplied supplemental content: various formats
    - *No standards, common practices, or even nomenclature*
- Content Identification and Linking
  - Use of persistent identifiers
    - *DOI is a terrific success*
- Versions and revisions
  - Differences between renditions (HTML, PDF, print, XML/SGML)
  - Identification of updates
    - *No standard practice regarding revisions and updates*
- Issue-level content for E-Journals
  - Covers, front matter, back matter
    - *Emerge best practice to include front & back in PDF*

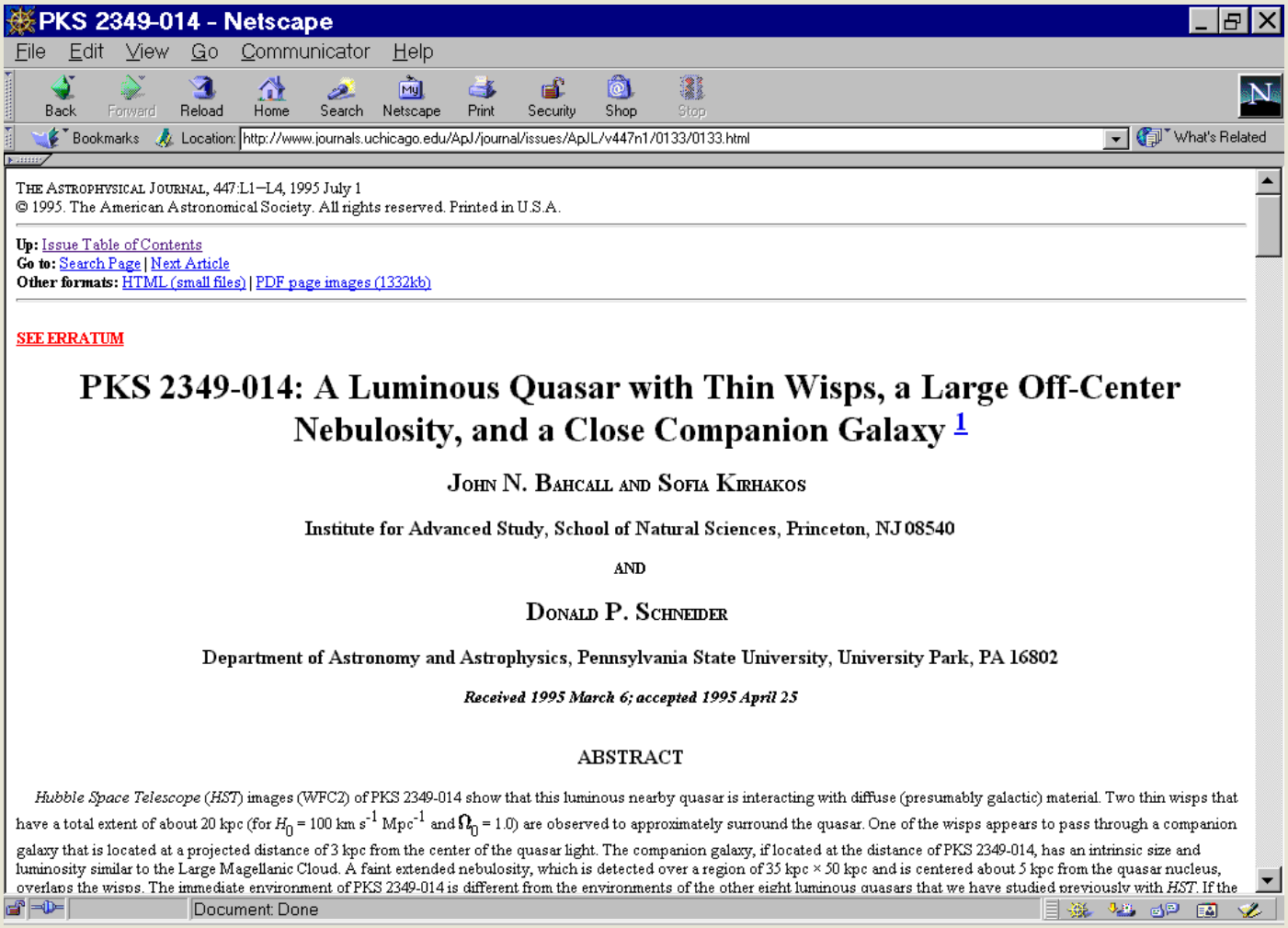
Portico's Preservation Approach for E-Journals

- Source file archiving
  - Preserve the components not the rendition
  - Include high-resolution files (PDF and figures) if available
  - All e-only components (data, media, etc.)
  - SGML / XML structured text by preference
    - HTML as last resort
- Preserve intellectual content not "look and feel" of HTML
  - HTML renditions are an artifact of current technology
    - Often dynamically generated
    - Fragile technology, overdue for change
- Preserve only essential features of the user interface
  - Reference linking, other content-based features
  - Not generic navigation or search or e-commerce features
- Why this approach?
  - Based on Mellon-funded study by Harvard University Library
  - Based on practical realities of works with multiple manifestations
  - Based on assessment as to instability of current web technologies

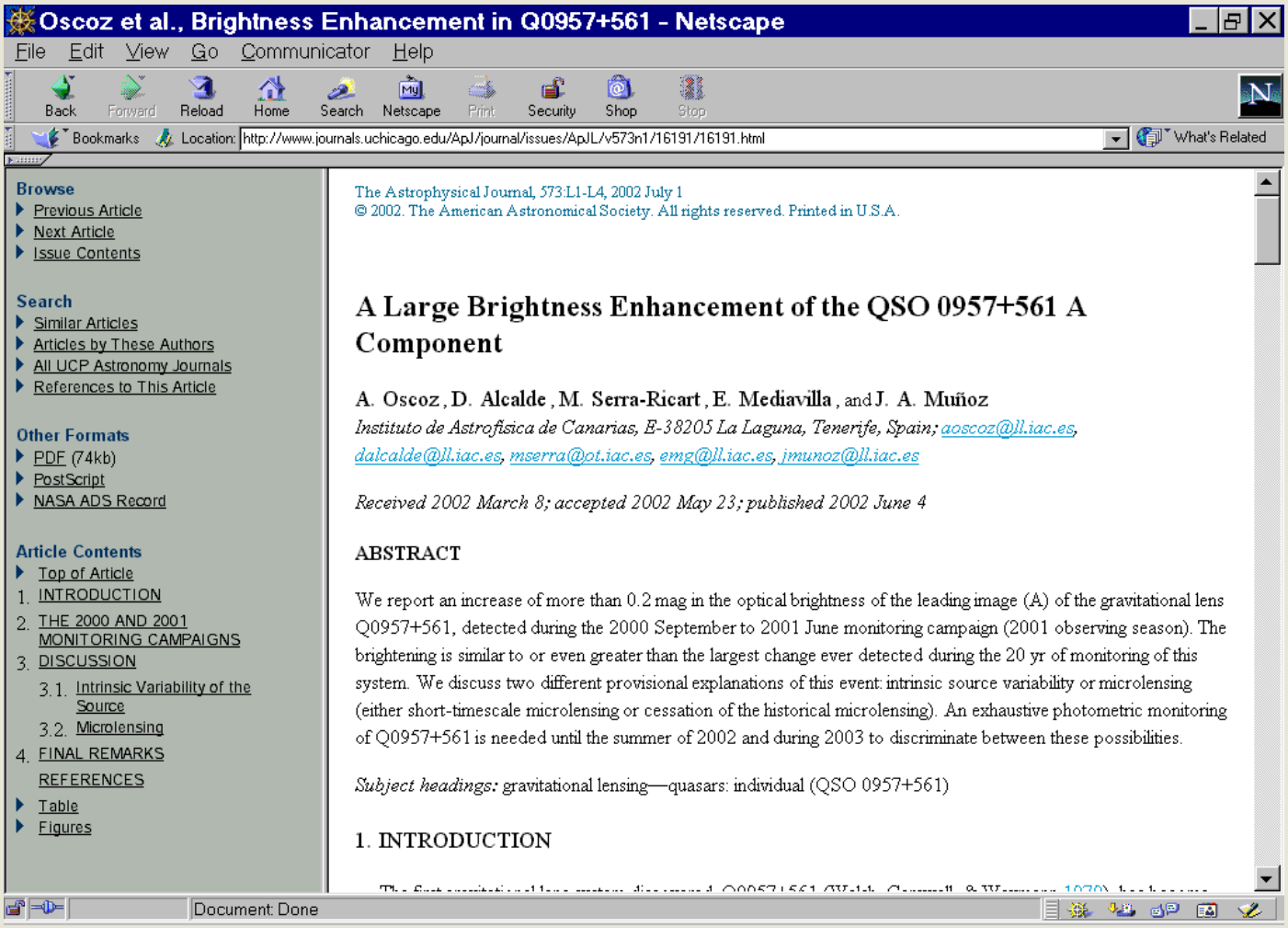
Astrophysical Journal, 1995 HTML on 2002 Browser



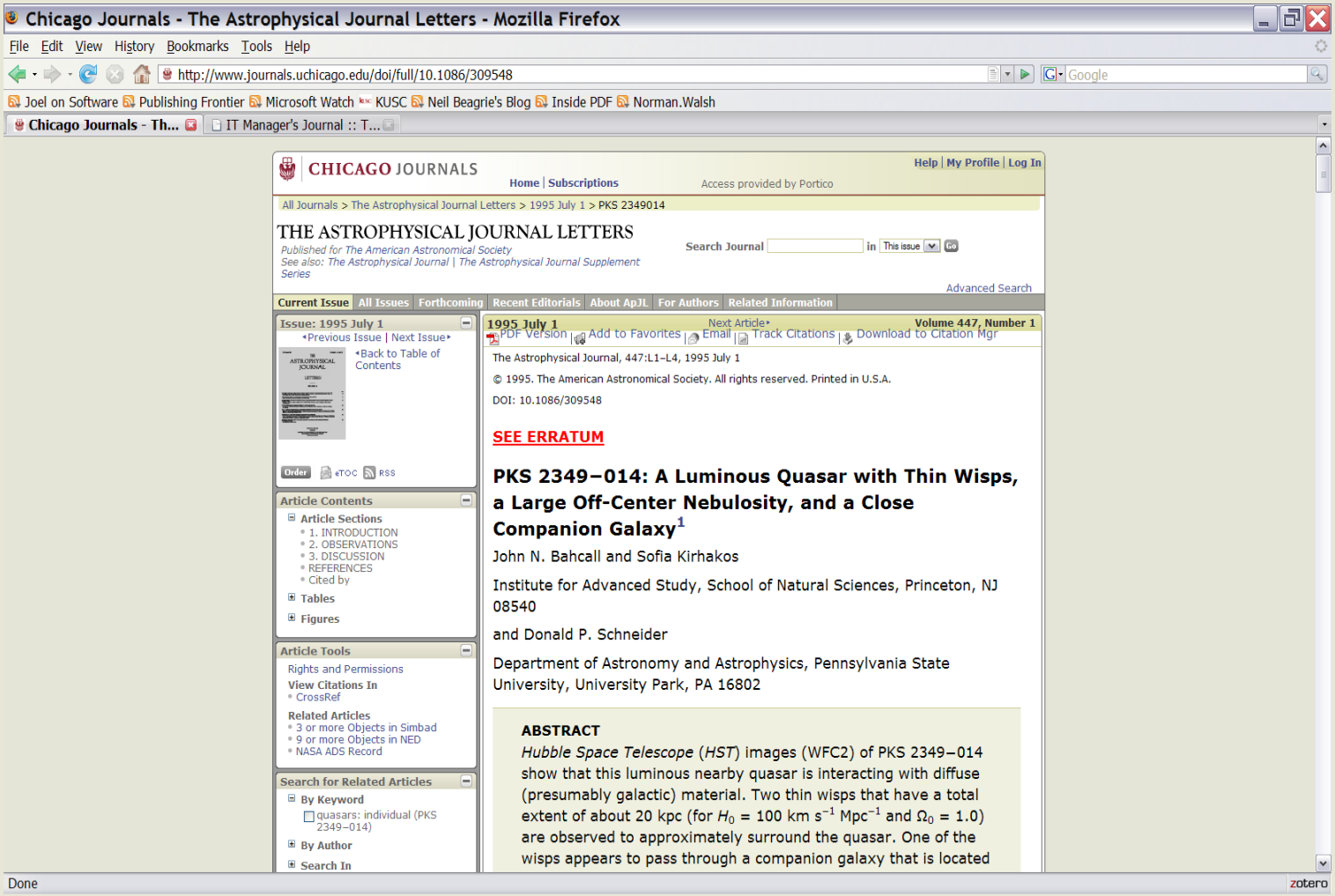
Astrophysical Journal, 1998 HTML on 2002 Browser



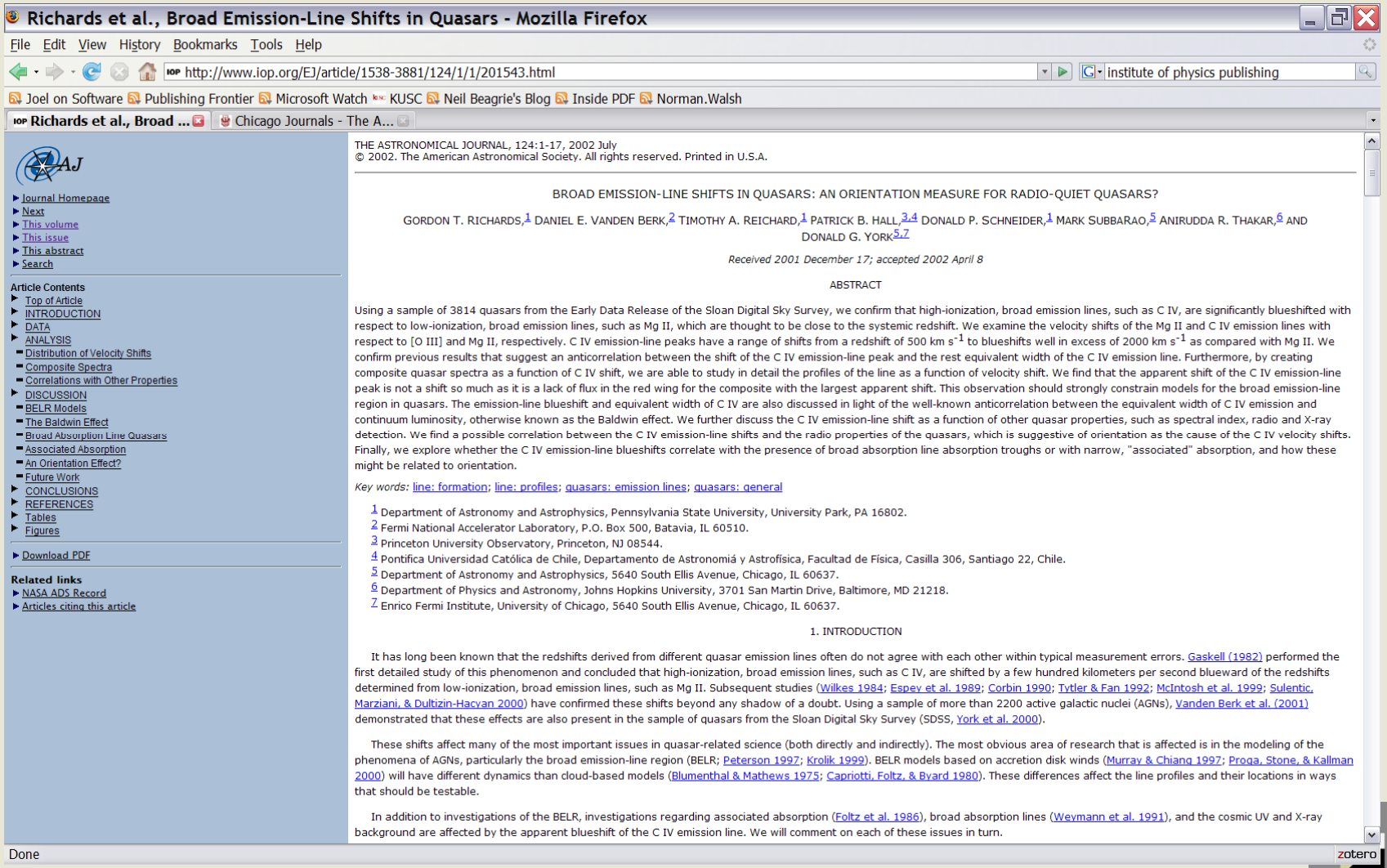
Astrophysical Journal, Framed HTML from 2002



Astrophysical Journal, 2007 HTML



Astronomical Journal, 2008 HTML



One Work: Multiple Renditions

- What Changed? Not much
- Exactly the same PDF in all cases
- Multiple HTML styles derived from SGML/XML source files
  - 5 major versions in 12 years
  - One hopes that web standards will stabilize eventually and publishers will stop changing the presentation so frequently
- Very Minor navigation changes
- Only significant functional change is additional reference linking
  - Made possible by the SGML/XML marked up full text
- Examples can be found from many other publishers as well

- What is the “work”? What is the object of preservation?
- An interesting philosophical and practical question

Portico Lessons Learned

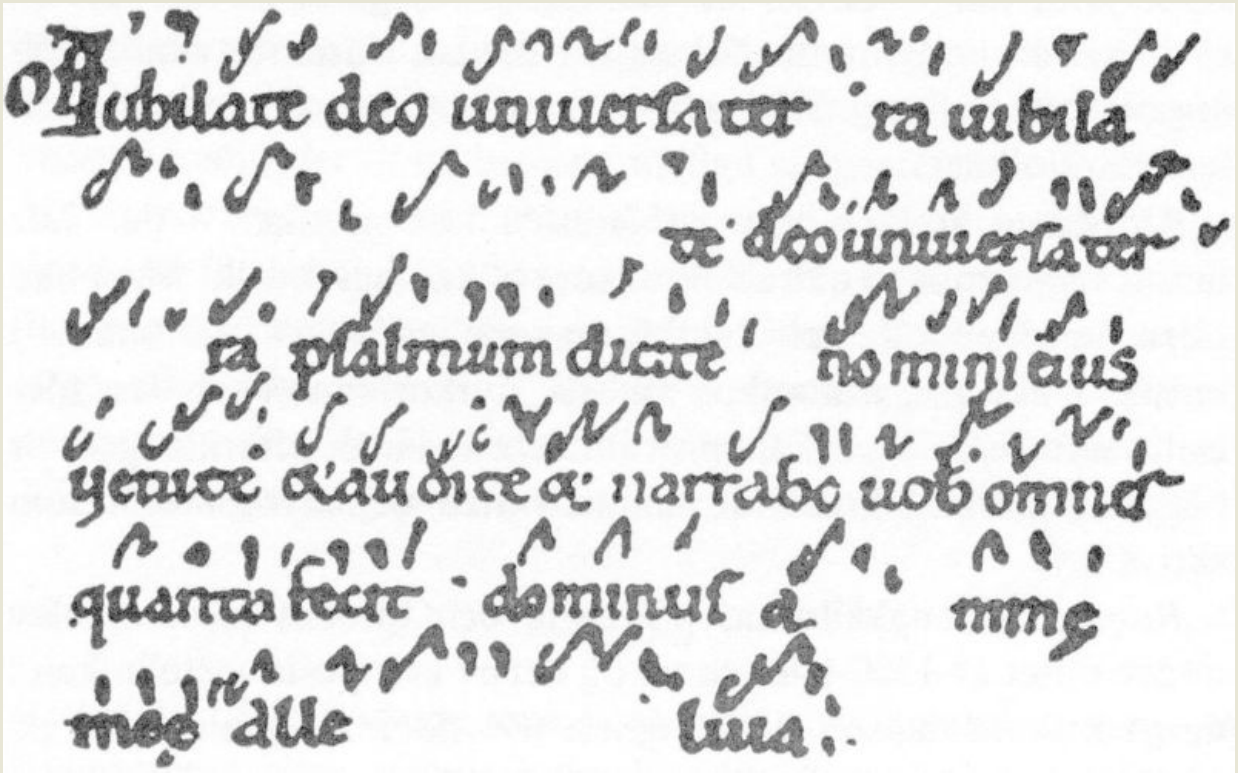
- Content is not perfect
- Software is not perfect
- People are not perfect
- Life is not perfect

- Audit trail is essential
  - Tool versions in event metadata helps trace problems
  - Trial by fire early in life of archive
  - CRL Audit report commended this feature of Portico systems

- Expect the unexpected!
  - Fixity check revealed 30 minute gap
  - Repaired from replicated content

EPILOGUE

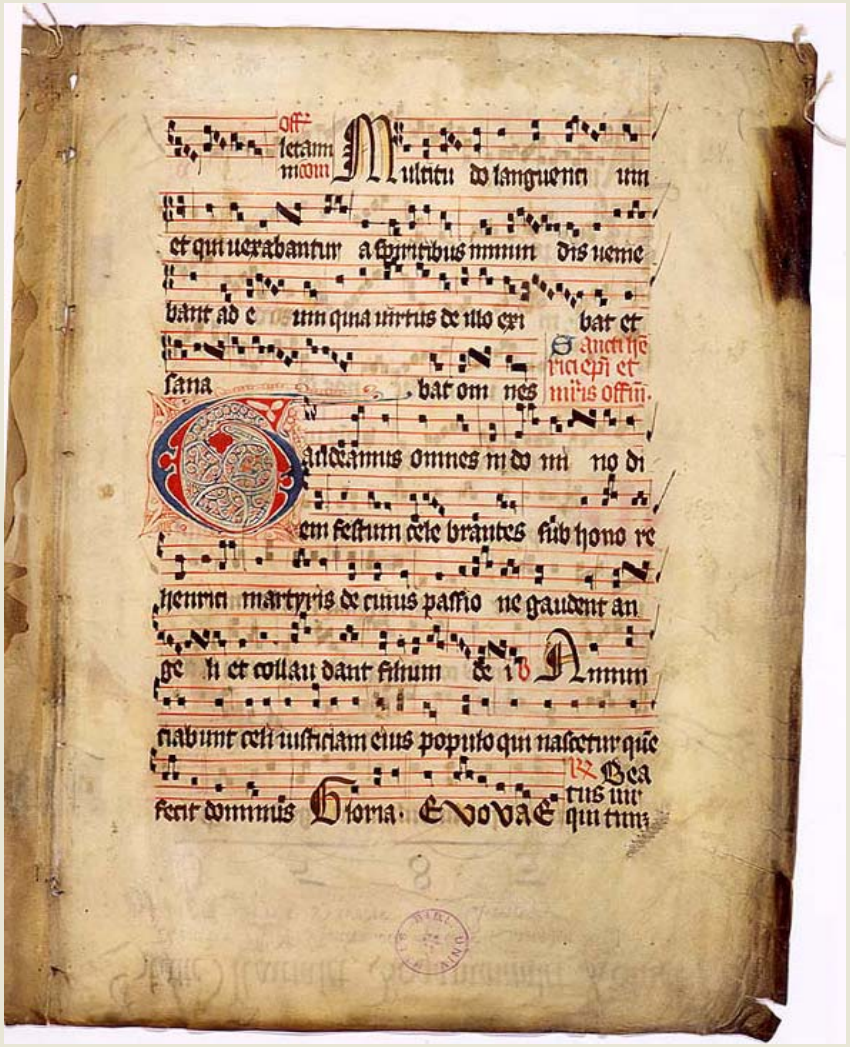
Some Musicological Analogies



"*Jubilate deo universa terra*" (9th Century)

Psalm verses in unheightened (staffless) neumes

From Wikipedia

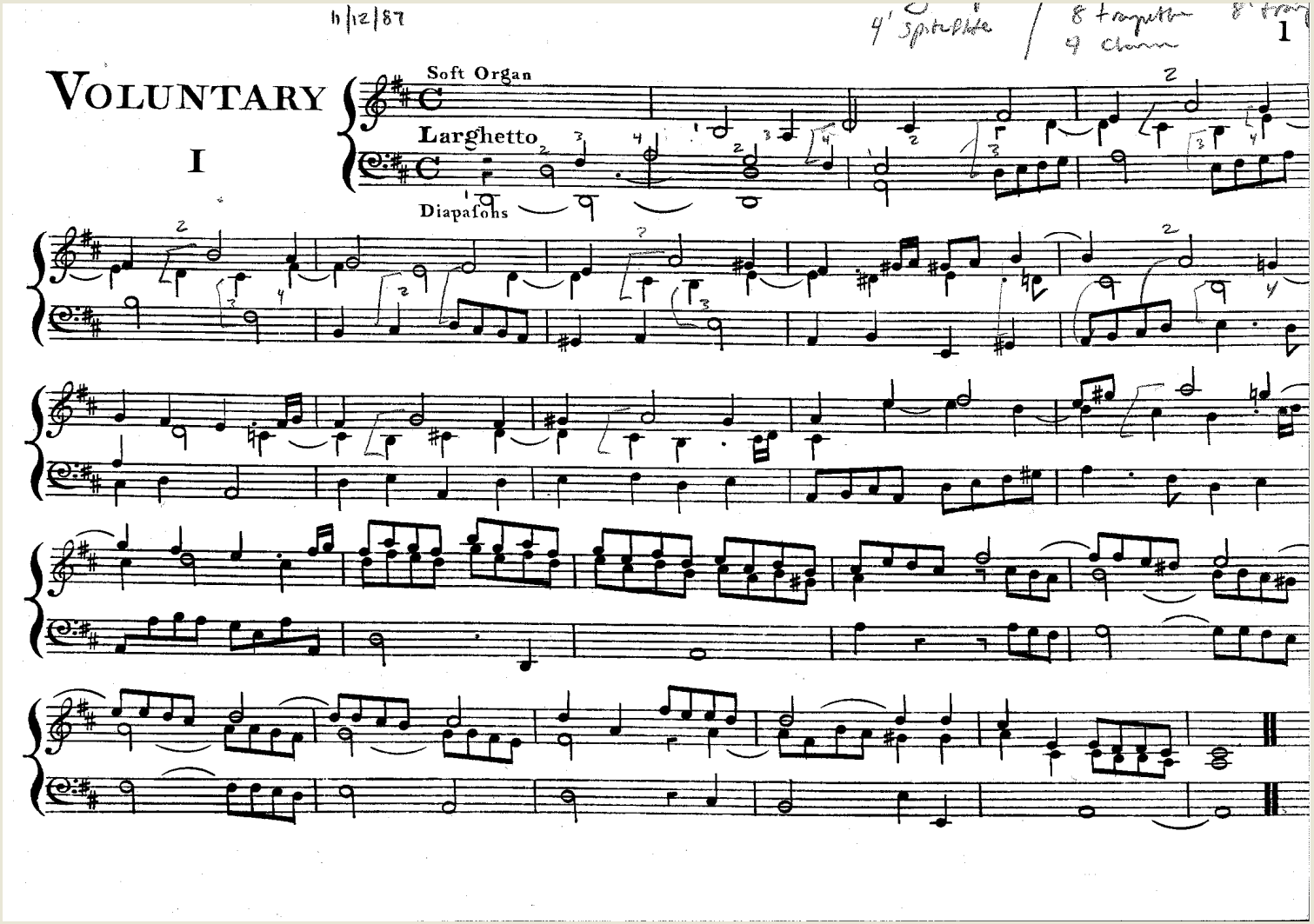


"*Gaudeamus omnes*"
from *Graduale Aboense*

Square notation

(14th-15th Century)

From Wikipedia



Organ Voluntary by William Boyce (18th Century)

Two Visions from the Past

- 18th-century sheet music printed on linen paper, as beautiful today as it was 250 years ago
- 19th-century sheet music printed on acidic paper, brittle and falling apart as the pages are turned

Technology doesn't always make things better

Which of these two visions will best describe our electronic content centuries from now?

Questions for the Future

- Will digital objects be lost forever?
  - Physical Preservation
- Will we be able to use them, interpret them, render them?
  - Logical (Format) Preservation
- Will they still have the appropriate information content?
  - Conceptual/Intellectual Preservation
- Were they the right objects in the first place?
  - Quality control

Digital preservation is very new. We need standards and best practices. We also need to hedge our bets through diversity.