Preserving the Scholarly Side of the Web

Adam Mikeal, Cody Green,
Alexey Maslov, Scott Phillips, John Leggett

Overview

1. Background
2. Challenges
3. Strategies
4. Case Study
5. Conclusion

Background

1. Background
2. Challenges
3. Strategies
4. Case Study
5. Conclusion

1

The Scholarly Web

- Information continues to migrate to the web
- Scholarly material is increasingly found on the web
- Lines between “digital libraries” and websites often blurred

Change

- Web content is ephemeral
- Web standards have high volatility

Preservation

- Stable methods of preservation of web content
- Available but inaccessible: *de facto* data loss

Challenges

1. Background
2. Challenges
3. Strategies
4. Case Study
5. Conclusion

2

Evolving Standards

- HTML: 5 different standards in 10 years
- Shift from SGML to XML
- Increasing separation of semantics and presentation

Lenient Validation

- Since MOSAIC, browsers have been lax
- “Standards” and “quirks” modes
- Not tenable in the long-term

Inconsistent Structure

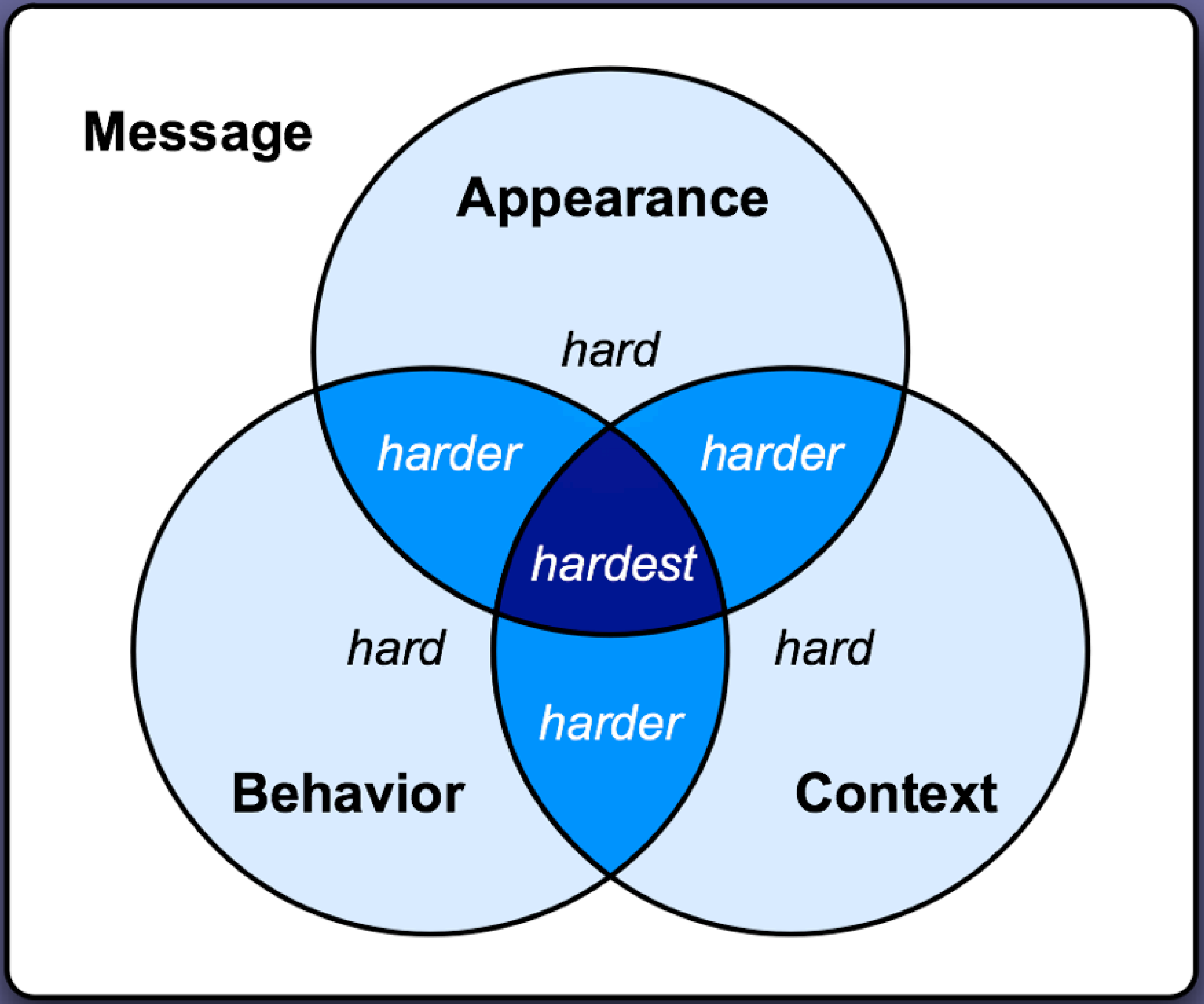
- Multiple authors, changing schema
- Ambiguous nomenclature

Message Preservation

- Inherent difficulties with definitions
- Document: physical manifestation of a message
- Levy: “bits of the material world... that we have imbued with the ability to speak”

3 Aspects of Document

- Appearance
- Behavior
- Context



3 Aspects of a Document

Author Intent

- Intersects with all 3 aspects
- Judging intent is frequently subjective
- Determined by others when author unavailable

Appearance

- For scholarly works, usually not as important
- Semantics/presentation separation clouds intent

Behavior

- Particularly important in digital environment
- Evolving web ecosystem shifts meaning of encodings
- Intrinsic message component or ubiquitous mechanism?

Context

- Both reader’s background and context at time of reading
- Focus on issues unique to the web: link networks
- Resilience to change in location and reference

Strategies

1. Background
2. Challenges
3. Strategies
4. Case Study
5. Conclusion

3

2 Strategies

- Emulation
- Migration

Emulation

- Document preserved as originally authored
- Interpreter emulates all original system functionality
- Interesting technical problem; frequently addressed

Emulation: Permutations

- Multiple standards, multiple interpretations
- Incompatible web ecosystems; each with separate interpreter
- Undocumented; originally intended ecosystem unknown

Emulation: Interpreter

- Sheer number of permutations very large
- Interpreter = browser
- Unreasonable to assume all permutations will be maintained

Emulation: Links

- For web content, link context is key
- Link semantics may evolve to create incompatibilities
- Technical and geo-political considerations

Emulation: Responsibility

- Preservation burden shifted to end user
- Better handled by trained archivists
- Hypertext archivists

Migration

- Documents periodically translated from older formats
- Access occurs with current tools
- Requires continual maintenance

Automatic Migration

- For large scale migration efforts, more cost efficient
- Predictability is high; results are consistent
- Potential exists for reuse

Automatic Migration

- Automatic methods unable to understand the *message*
- No consideration of author intent
- Necessary pre-conditions; data consistency

Manual Migration

- Significant human intervention during process
- Potential for inconsistency and transcription error
- Higher cost per document; not tenable on large collections

Manual Migration

- Greater flexibility with inconsistent data, lenient validation
- Able to interpret author intent
- Understanding of document *message*

Link Migration

- Methods of preservation depend on degree of control
- 2 link components: pointer and target
- 3 link types: *internal*, *in-coming*, *out-going*

Internal Links

- Both pointer and target under control of same archivist
- Preservation through simultaneous change to both components
- Change in target location triggers modification to all associated pointers

In-coming Links

- Target under control of the archivist, but pointer is not
- To preserve the validity of the link, original target location must be maintained
- Transient location mechanisms add layer of indirection

Out-going Links

- Pointer under control of the archivist, but target resource is not
- Target likely to change or disappear
- Point to original location anyway; point to an archived copy

Case Study

1. Background
2. Challenges
3. Strategies
4. Case Study
5. Conclusion

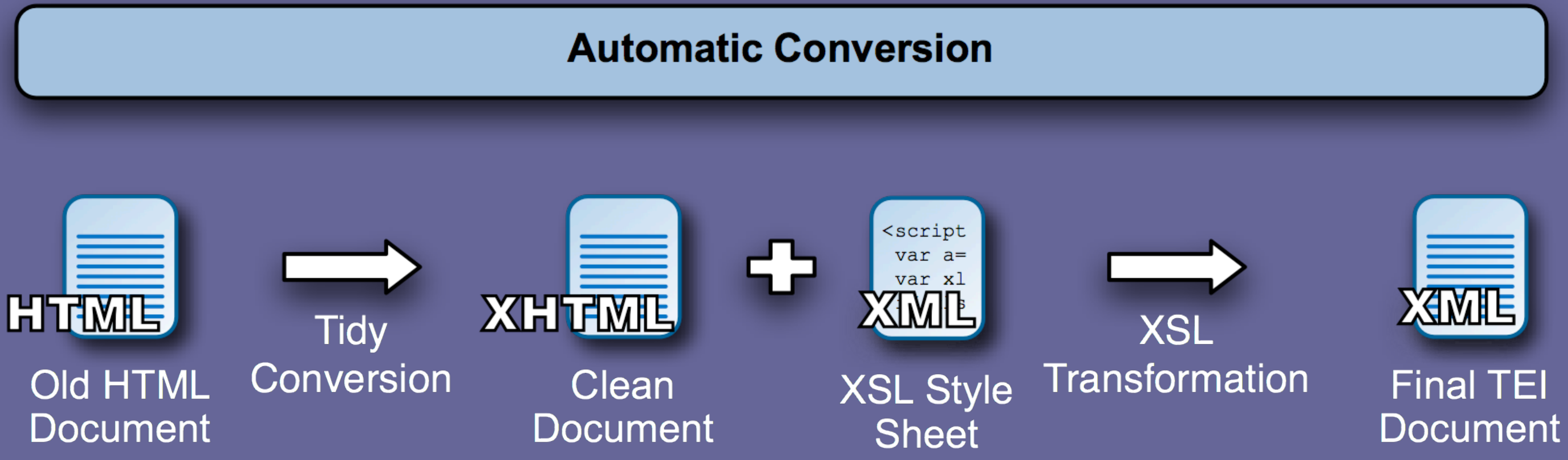
4

Journal of Digital Information

- Peer-reviewed, electronic-only journal
- First published in 1997
- Older, non-standard technologies

Goals

- Immediate: upgrade to OJS
- Short-term: improved services
- Long-term: stable preservation format



Automatic Conversion

Old HTML Document

Tidy Conversion

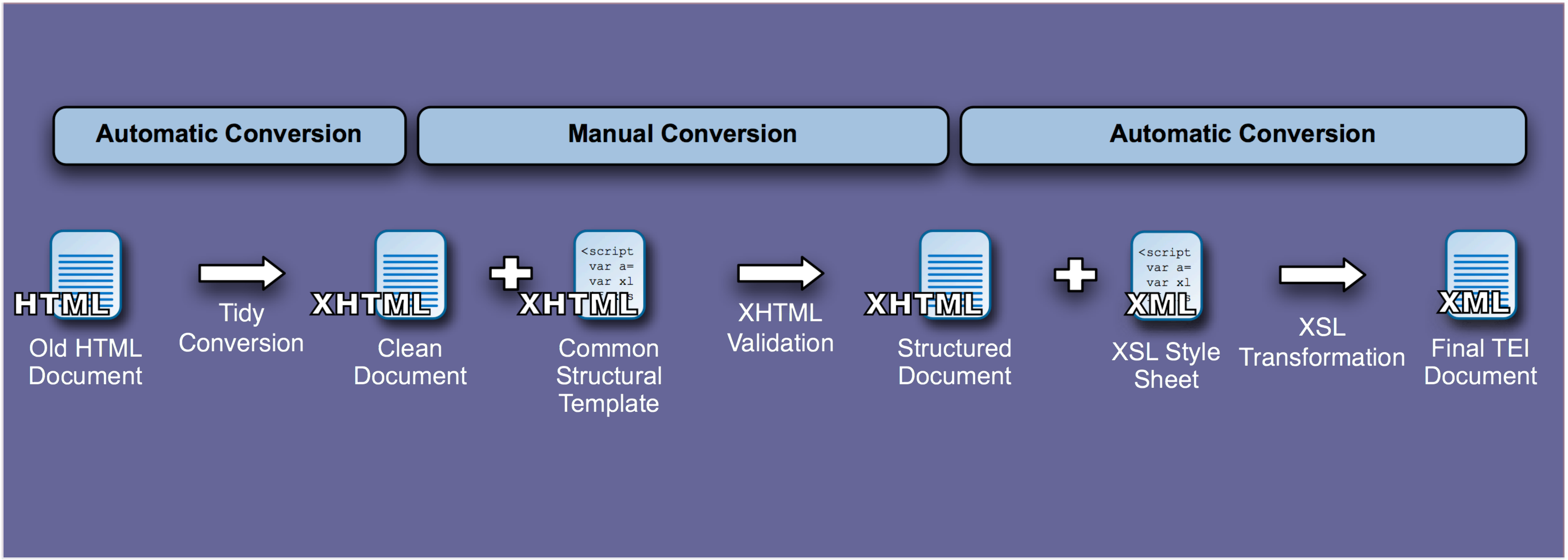
Clean Document

XSL Style Sheet

XSL Transformation

Final TEI Document

- Inconsistent structure
- Evolving standards
- Message preservation (back to author intent)



- Multi-stage process
- Manual stage allowed decision on intent
- Message preservation, translation

Link Migration

- Internal: manual change to pointer, target
- In-coming: two-phased solution; preserve targets, redirects
- Out-going: left as originally authored; context issues

Conclusion

1. Background
2. Challenges
3. Strategies
4. Case Study
5. Conclusion

5

Lessons Learned

- Automation useful and necessary
- Significant issues create difficulties:
  structural inconsistency, author intent
- Issues should be addressed at document creation

Future Work

- Creation of conventions for document structure
- Deprecation of “quirks” mode
- Better automation tools
- Recognition of new role: the hypertext archivist

Call to Arms

- Issue needs attention from digital preservation community
- Web is full of documents in nearly obsolete formats
- Effort now to preserve this portion of the scholarly record
- Effort directed at mechanisms to prevent perpetuating cycle