PORTICO

# MESSES AND MEANINGS

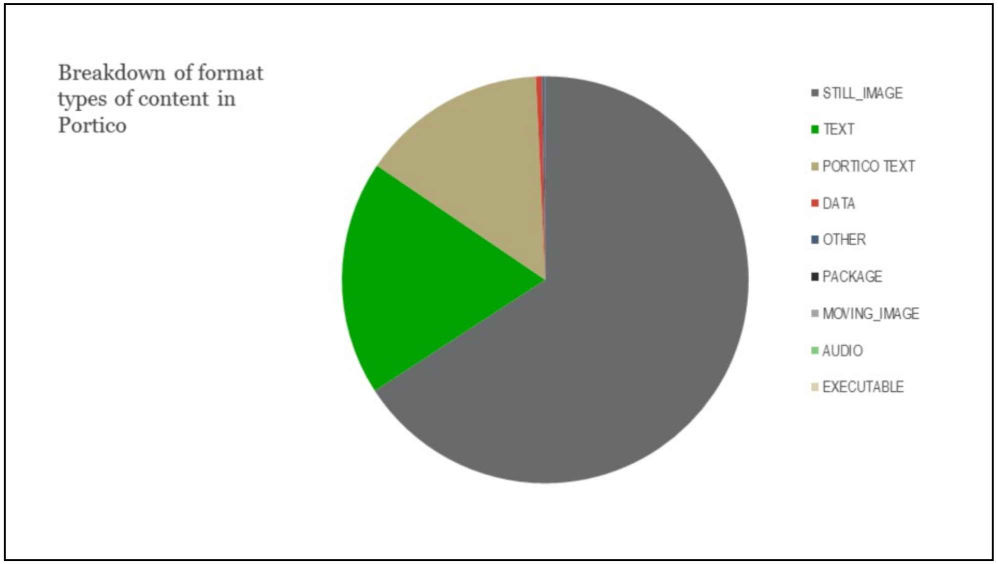
Sheila Morrissey
*@sheilaMorr*

DPC Re:Format - What is file format obsolescence and does it really exist?
York
22 June 2016

PORTICO

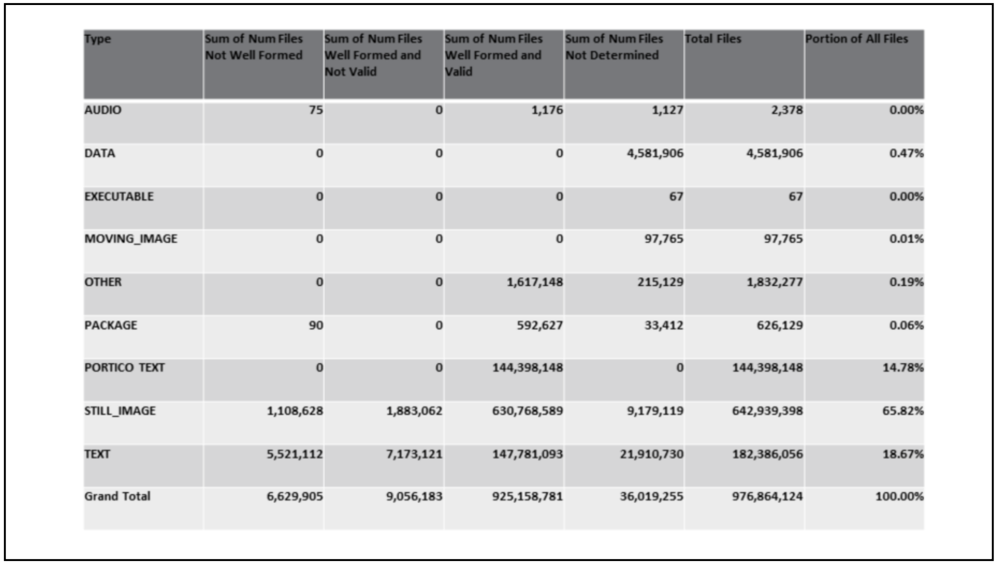
PORTICO

Portico is a preservation service for digital publications, including electronic journals, books, and historical collections.

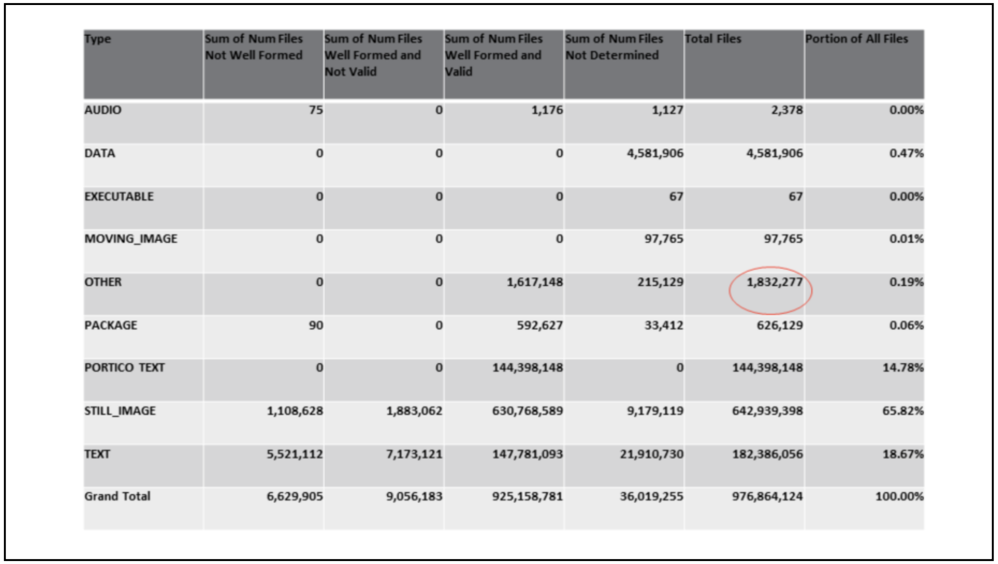


Text: **pdf 72million** 7.39 postscriptXML, SGML, Javascript, latex, perl,utf-8, css, csv, html(.3), asm, C

| | | | |
|---|---|---|---|
| application/vnd.corel-draw (cdr) | 13 | 0.00% | |
| application/x-ptc-paintshoppro (psp) | 1 | 0.00% | |
| image/gif (gif) | | 10,771 | 39.32% |
| **image/jpeg (jpg)** | **149,779,348** | **15.33**% | |
| image/png (png) | 9,170,228 | 0.94% | |
| **image/tiff (tif)** | **99,870,761** | **10.22%** | |
| image/vnd.adobe.photoshop (psd) | 3,744 | 0.00% | |
| image/x-ms-bmp (bmp) | 11 | | 0.00% |
| image/x-ms-bmp-v3 (bmp) | 4,836 | | 0.00% |
| image/x-wmf (wmf) | 286 | | 0.00% |

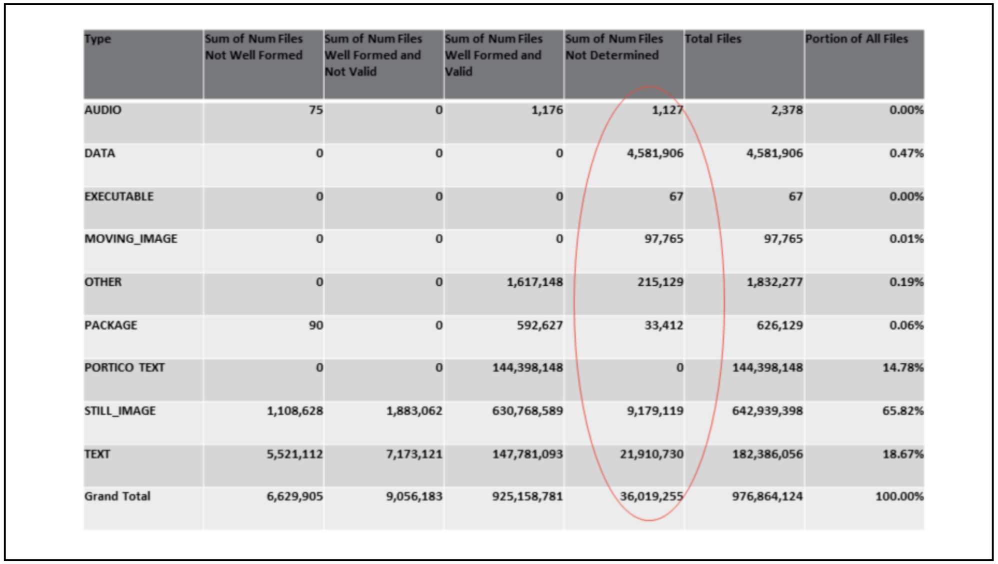


| Type | Sum of Num Files Not Well Formed | Sum of Num Files Well Formed and Not Valid | Sum of Num Files Well Formed and Valid | Sum of Num Files Not Determined | Total Files | Portion of All Files |
|---|---|---|---|---|---|---|
| AUDIO | 75 | 0 | 1,176 | 1,127 | 2,378 | 0.00% |
| DATA | 0 | 0 | 0 | 4,581,906 | 4,581,906 | 0.47% |
| EXECUTABLE | 0 | 0 | 0 | 67 | 67 | 0.00% |
| MOVING_IMAGE | 0 | 0 | 0 | 97,765 | 97,765 | 0.01% |
| OTHER | 0 | 0 | 1,617,148 | 215,129 | 1,832,277 | 0.19% |
| PACKAGE | 90 | 0 | 592,627 | 33,412 | 626,129 | 0.06% |
| PORTICO TEXT | 0 | 0 | 144,398,148 | 0 | 144,398,148 | 14.78% |
| STILL_IMAGE | 1,108,628 | 1,883,062 | 630,768,589 | 9,179,119 | 642,939,398 | 65.82% |
| TEXT | 5,521,112 | 7,173,121 | 147,781,093 | 21,910,730 | 182,386,056 | 18.67% |
| Grand Total | 6,629,905 | 9,056,183 | 925,158,781 | 36,019,255 | 976,864,124 | 100.00% |



| Type | Sum of Num Files Not Well Formed | Sum of Num Files Well Formed and Not Valid | Sum of Num Files Well Formed and Valid | Sum of Num Files Not Determined | Total Files | Portion of All Files |
|---|---|---|---|---|---|---|
| AUDIO | 75 | 0 | 1,176 | 1,127 | 2,378 | 0.00% |
| DATA | 0 | 0 | 0 | 4,581,906 | 4,581,906 | 0.47% |
| EXECUTABLE | 0 | 0 | 0 | 67 | 67 | 0.00% |
| MOVING_IMAGE | 0 | 0 | 0 | 97,765 | 97,765 | 0.01% |
| OTHER | 0 | 0 | 1,617,148 | 215,129 | 1,832,277 | 0.19% |
| PACKAGE | 90 | 0 | 592,627 | 33,412 | 626,129 | 0.06% |
| PORTICO TEXT | 0 | 0 | 144,398,148 | 0 | 144,398,148 | 14.78% |
| STILL_IMAGE | 1,108,628 | 1,883,062 | 630,768,589 | 9,179,119 | 642,939,398 | 65.82% |
| TEXT | 5,521,112 | 7,173,121 | 147,781,093 | 21,910,730 | 182,386,056 | 18.67% |
| Grand Total | 6,629,905 | 9,056,183 | 925,158,781 | 36,019,255 | 976,864,124 | 100.00% |

Unknown/bitstream  not identified



| Type | Sum of Num Files Not Well Formed | Sum of Num Files Well Formed and Not Valid | Sum of Num Files Well Formed and Valid | Sum of Num Files Not Determined | Total Files | Portion of All Files |
|---|---|---|---|---|---|---|
| AUDIO | 75 | 0 | 1,176 | 1,127 | 2,378 | 0.00% |
| DATA | 0 | 0 | 0 | 4,581,906 | 4,581,906 | 0.47% |
| EXECUTABLE | 0 | 0 | 0 | 67 | 67 | 0.00% |
| MOVING_IMAGE | 0 | 0 | 0 | 97,765 | 97,765 | 0.01% |
| OTHER | 0 | 0 | 1,617,148 | 215,129 | 1,832,277 | 0.19% |
| PACKAGE | 90 | 0 | 592,627 | 33,412 | 626,129 | 0.06% |
| PORTICO TEXT | 0 | 0 | 144,398,148 | 0 | 144,398,148 | 14.78% |
| STILL_IMAGE | 1,108,628 | 1,883,062 | 630,768,589 | 9,179,119 | 642,939,398 | 65.82% |
| TEXT | 5,521,112 | 7,173,121 | 147,781,093 | 21,910,730 | 182,386,056 | 18.67% |
| Grand Total | 6,629,905 | 9,056,183 | 925,158,781 | 36,019,255 | 976,864,124 | 100.00% |

**Undetermined**

**Audio: mpeg (undetermined)**

**Data**:

fits, maple mathematica,rtf, mathcad, stata data and programs

pdb chem cdx/cif/jmol-voxel,mol,(80K)

word 527,142

excel 136,335

ppt 141,006

open office 390K

**Executable**

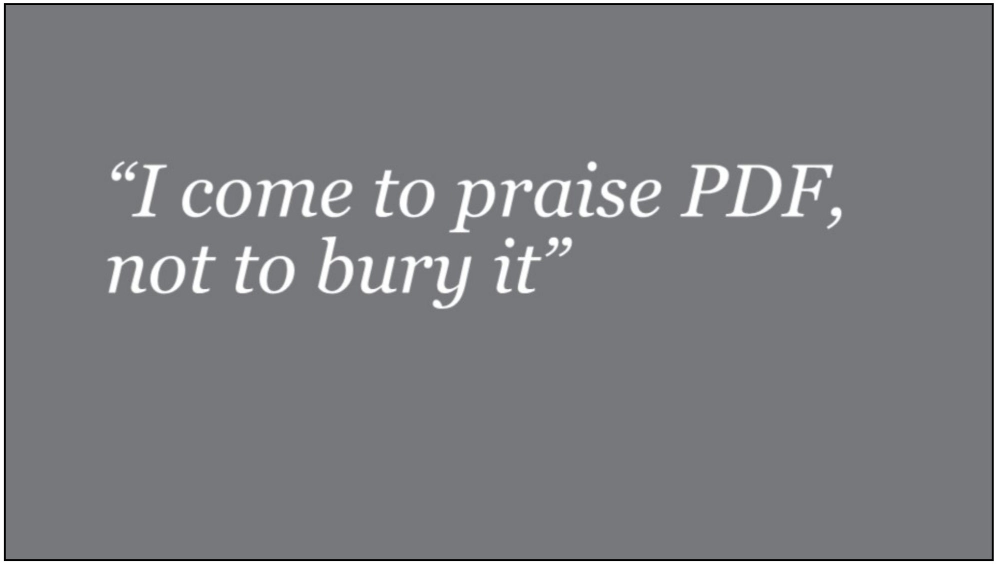
**Moving Image**
**realmedia, flash, avi, mp4, mpeg, quicktime, wmv**

**Still-image** not determined:  almost entirely PNG  1.4 % of our image files

**text** not determined dataset.toc, dataset.xml

**Package**:  epub, stuffit

OTHER Bytestream:  not identified



*"I come to praise PDF, not to bury it"*

## *Well, maybe just a little…*

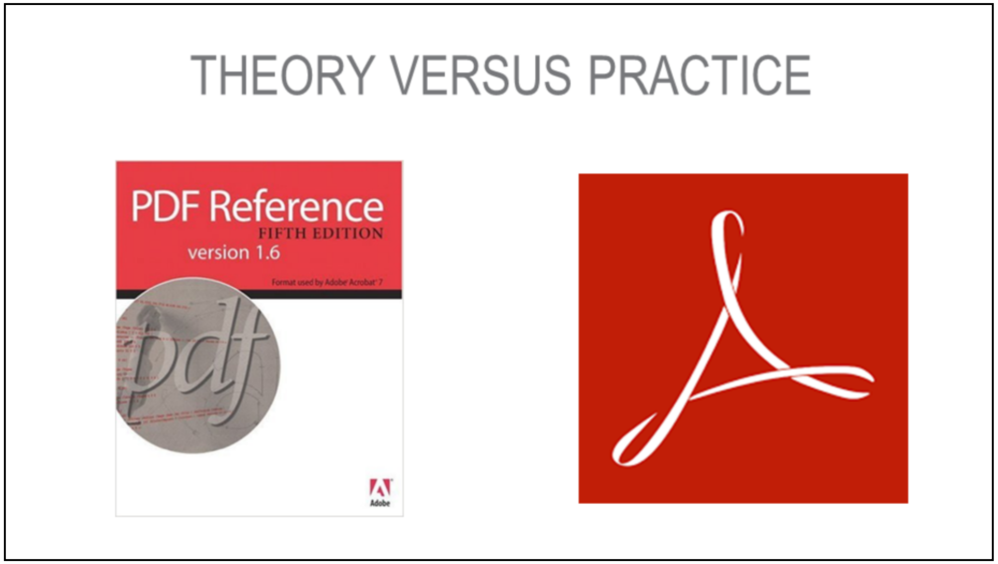
# OMG-WTF-PDF

[PDF Ambiguity and Obfuscation]

Julia Wolf
FireEye

2010 March 31
Troopers 11

https://www.troopers.de/wp-content/uploads/2011/04/TR11_Wolf_OMG_PDF.pdf



THEORY VERSUS PRACTICE

1248 pages

Calvin is a cataract, a primeval forest, a demonic power, something directly down from Himalaya, absolutely Chinese, strange, mythological; I lack completely the means, the suction cups, even to assimilate this phenomenon, not to speak of presenting it adequately.

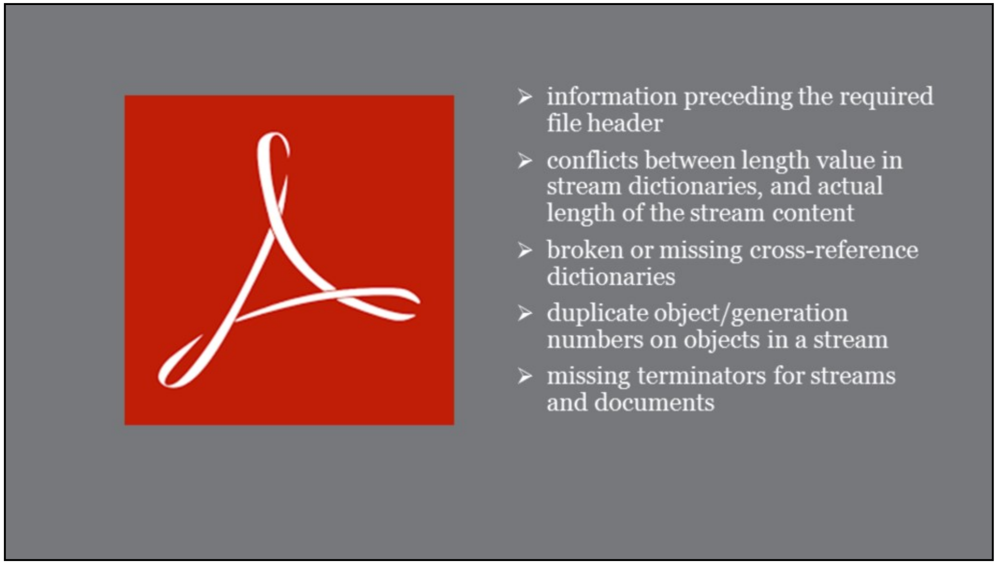
-Karl Barth, *Revolutionary Theology in the Making: Barth-Thurneysen Correspondence 1914-1925*, trans. James D. Smart (Richmond: John Knox Press, 1964), 101.

Incredible accomplishment  23 years ago 2013

Interoperable (before Java)

“Interchange PostScript”  Analogous to simplification from SGML to XML

Portable document format

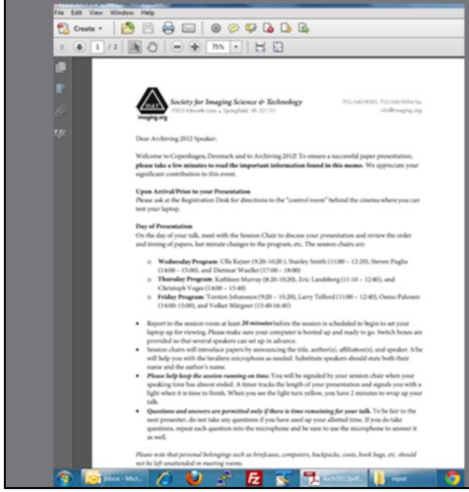


- information preceding the required file header
- conflicts between length value in stream dictionaries, and actual length of the stream content
- broken or missing cross-reference dictionaries
- duplicate object/generation numbers on objects in a stream
- missing terminators for streams and documents

- (for example, files produced on older versions of Apple’s operating system, which include Apple Single and Apple Double encoding information preceding the header)

Other developers of PDF applications clearly have struggled with the gap between PDF “in theory” and PDF instances in actual practice. Comments in the PDFParser source code from Adobe’s PDFBox Java application note some of the various compensatory actions taken silently to repair non-conforming documents.[12] The widely-used iText library includes an “isRebuilt()” method to indicate that it has “repaired” syntactic errors

# THEORY VERSUS PRACTICE



Society for Imaging Science & Technology
7003 Kilworth Lane • Springfield, VA 22151
imaging.org

703/642-9090, 703/642-9094 fax
info@imaging.org

Dear Archiving 2012 Speaker:

Welcome to Copenhagen, Denmark and to Archiving 2012! To ensure a successful paper presentation, **please take a few minutes to read the important information found in this memo.** We appreciate your significant contribution to this event.

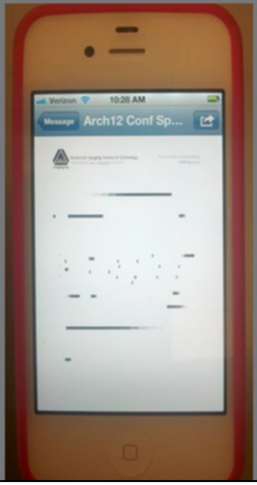
**Upon Arrival/Prior to your Presentation**
Please ask at the Registration Desk for directions to the "control room" behind the cinema where you can test your laptop.

**Day of Presentation**
On the day of your talk, meet with the Session Chair to discuss your presentation and review the order and timing of papers, last minute changes to the program, etc. The session chairs are:

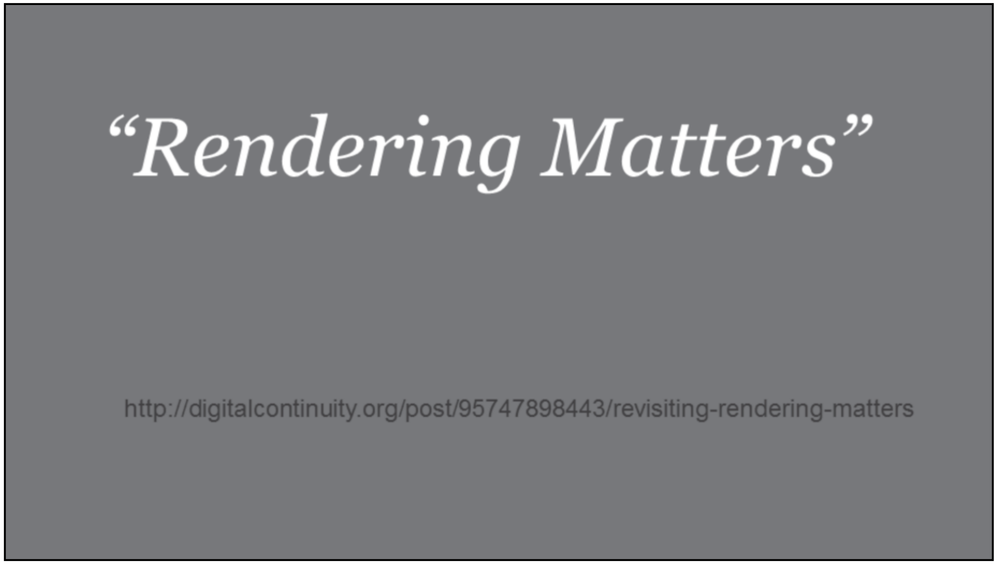
- ○ **Wednesday Program**: Ulla Kejser (9:20-10:20 ), Stanley Smith (11:00 – 12:20), Steven Puglia (14:00 – 15:00), and Dietmar Wueller (17:00 – 18:00)
- ○ **Thursday Program**: Kathleen Murray (8:20-10:20), Eric Landsberg (11:10 – 12:40), and Christoph Voges (14:00 – 15:40)
- ○ **Friday Program**: Torsten Johansson (9:20 – 10:20), Larry Telford (11:00 – 12:40), Osmo Palonen (14:00-15:00), and Volker Märgner (15:40-16:40)

- Report to the session room at least ***20 minutes*** before the session is scheduled to begin to set your laptop up for viewing. Please make sure your computer is booted up and ready to go. Switch boxes are provided so that several speakers can set up in advance.
- Session chairs will introduce papers by announcing the title, author(s), affiliation(s), and speaker. S/he will help you with the lavaliere microphone as needed. Substitute speakers should state both their name and the author's name.
- ***Please help keep the session running on time.*** You will be signaled by your session chair when your speaking time has almost ended. A timer tracks the length of your presentation and signals you with a light when it is time to finish. When you see the light turn yellow, you have 2 minutes to wrap up your talk.
- ***Questions and answers are permitted only if there is time remaining for your talk.*** To be fair to the next presenter, do not take any questions if you have used up your allotted time. If you do take questions, repeat each question into the microphone and be sure to use the microphone to answer it as well.

*Please note that personal belongings such as briefcases, computers, backpacks, coats, book bags, etc. should not be left unattended in meeting rooms.*



Jhove: WFV



“Rendering Matters”

http://digitalcontinuity.org/post/95747898443/revisiting-rendering-matters

Tools break – iText, PDFBox – we just chain them split PDF files

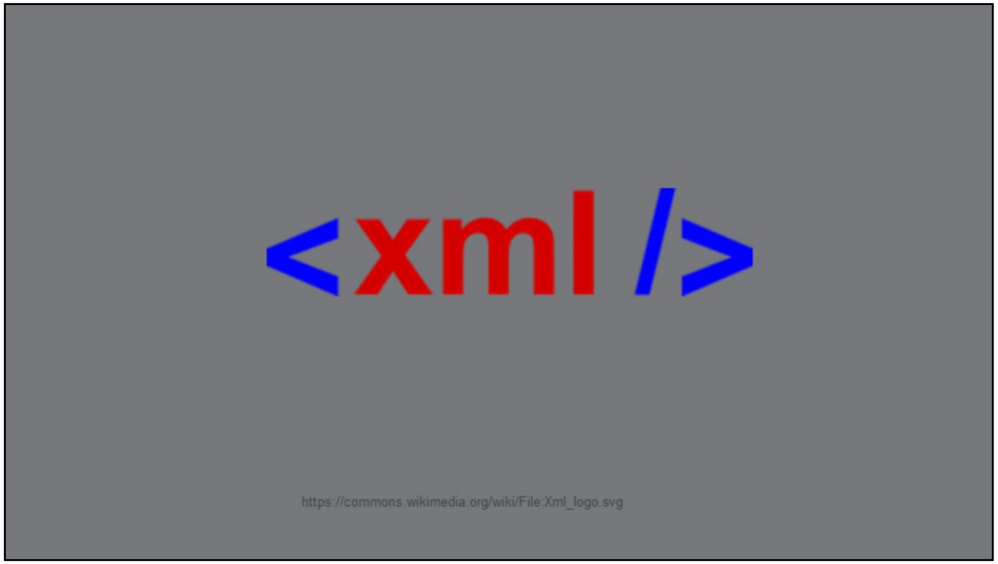
## CHARACTERISTICS OF APPLICATIONS AS WELL AS OF INSTANCES

- ➢ What is “standard”?
- ➢ What is “forgiven”?
- ➢ What is “repaired”?
- ➢ What is broken?

Characterize the renderers

VeraPDF better expression of failures of conformance
Duff Johnson, PDF Association – DPC, OPF

Match it up to render characterizer

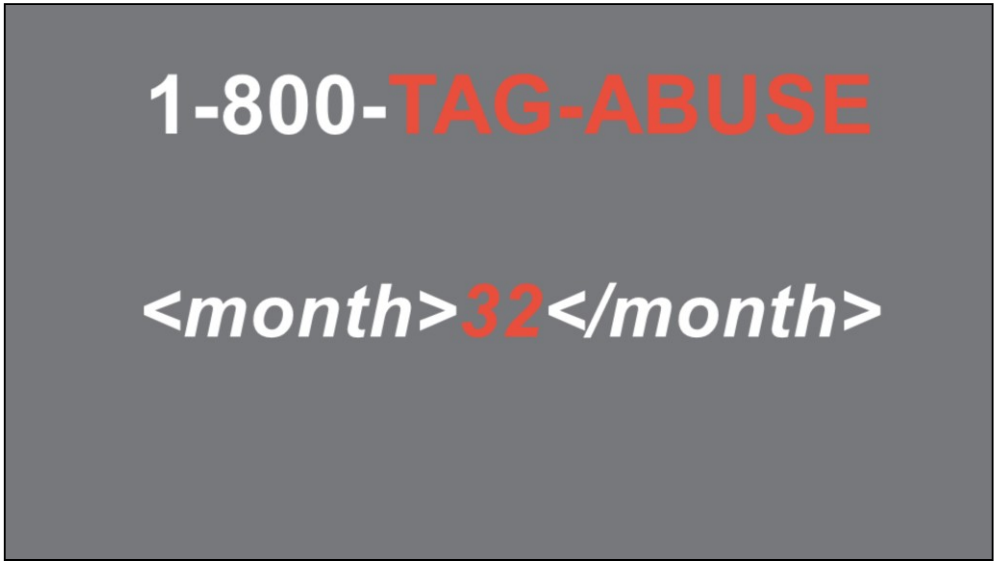


https://commons.wikimedia.org/wiki/File:Xml_logo.svg

Interoperable

Never validated  - NWF, WFNV, Encoding (easy stuff - -syntax)

Trickier:  procedural semantics  (Display italic vs regular,), “Implied text” - -boiler plate –good practice – hard for preservation (encapsulate)



1-800-TAG-ABUSE

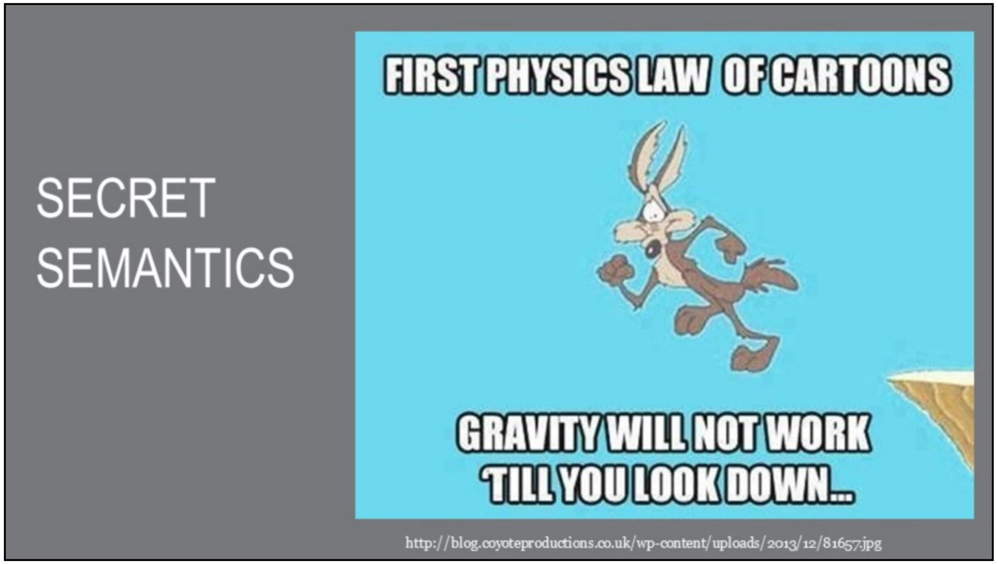
<month>32</month>

Month 32

Impedance mismatch  -- fine/coarse grain  Synonymy

It’s a tree

Sell-by date  -- sophisticated use

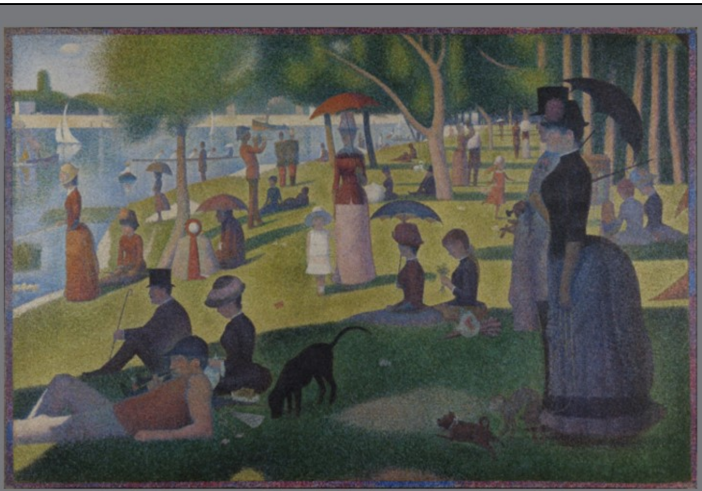


SECRET SEMANTICS

FIRST PHYSICS LAW OF CARTOONS
GRAVITY WILL NOT WORK 'TILL YOU LOOK DOWN...

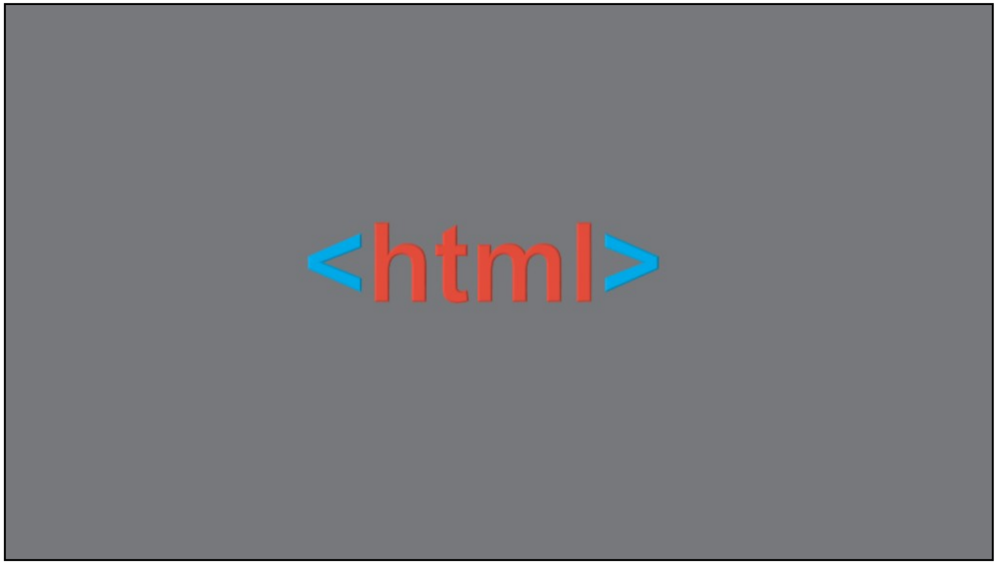
http://blog.coyoteproductions.co.uk/wp-content/uploads/2013/12/81657.jpg

Separation of concerns -



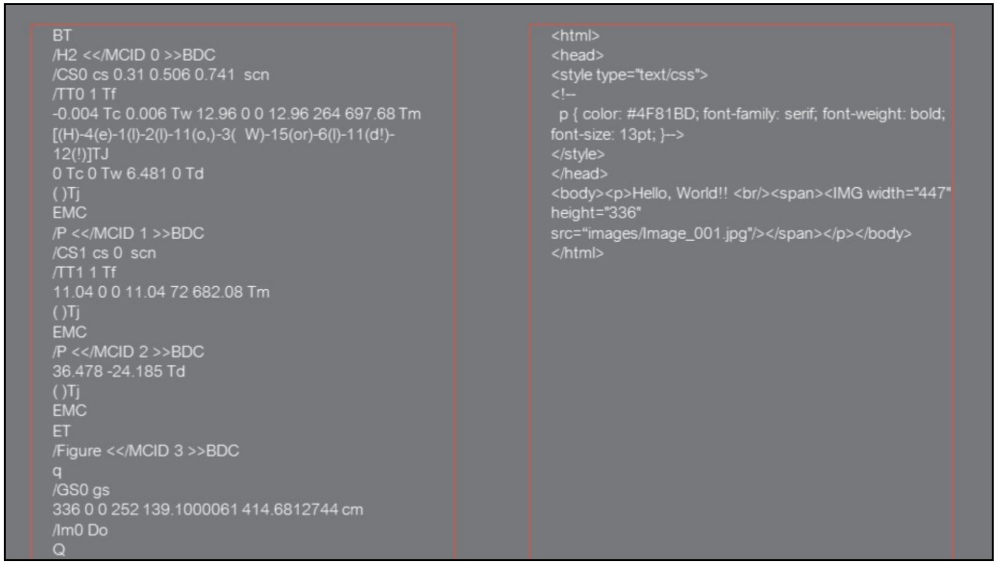
A Sunday Afternoon on the Island of La Grande Jatte  Georges Seurat - A Sunday on La Grande Jatte -- 1884 - Google Art Project

https://www.google.com/culturalinstitute/beta/asset/twGyqq52R-lYpA



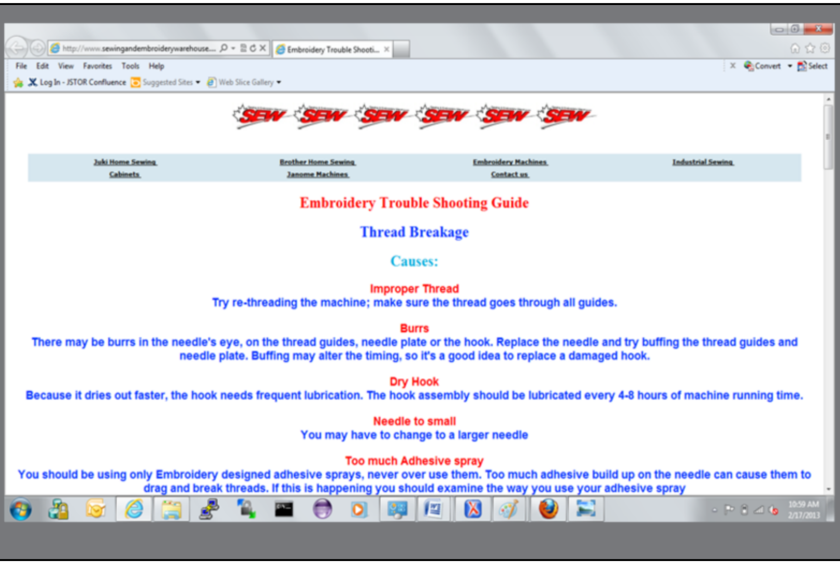
<html>

## Hello, World!!



```
BT
/H2 <</MCID 0 >>BDC
/CS0 cs 0.31 0.506 0.741  scn
/TT0 1 Tf
-0.004 Tc 0.006 Tw 12.96 0 0 12.96 264 697.68 Tm
[(H)-4(e)-1(l)-2(l)-11(o,)-3(  W)-15(or)-6(l)-11(d!)-
12(!)]TJ
0 Tc 0 Tw 6.481 0 Td
( )Tj
EMC
/P <</MCID 1 >>BDC
/CS1 cs 0  scn
/TT1 1 Tf
11.04 0 0 11.04 72 682.08 Tm
( )Tj
EMC
/P <</MCID 2 >>BDC
36.478 -24.185 Td
( )Tj
EMC
ET
/Figure <</MCID 3 >>BDC
q
/GS0 gs
336 0 0 252 139.1000061 414.6812744 cm
/Im0 Do
Q
```

```
<html>
<head>
<style type="text/css">
<!–
  p { color: #4F81BD; font-family: serif; font-weight: bold;
font-size: 13pt; }–>
</style>
</head>
<body><p>Hello, World!! <br/><span><IMG width="447"
height="336"
src="images/Image_001.jpg"/></span></p></body>
</html>
```



Juki Home Sewing | Brother Home Sewing | Embroidery Machines | Industrial Sewing
Cabinets | Janome Machines | Contact us

## Embroidery Trouble Shooting Guide

## Thread Breakage

### Causes:

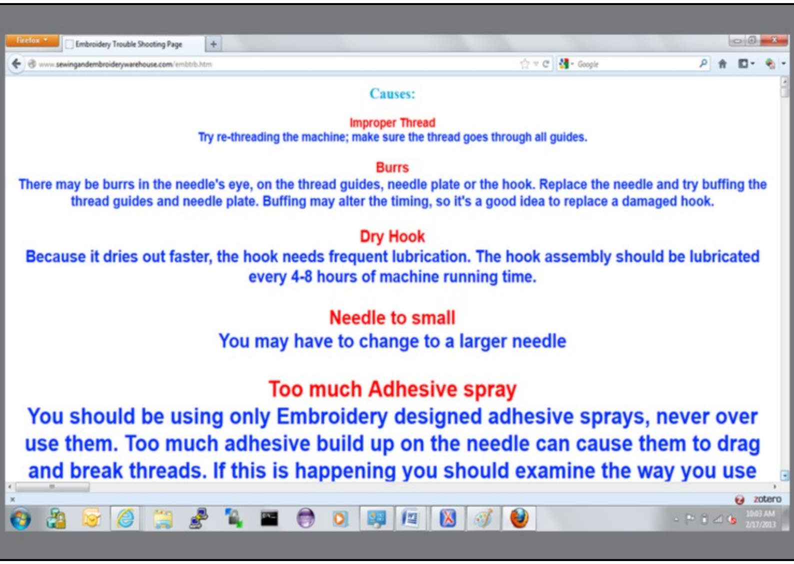
**Improper Thread**
Try re-threading the machine; make sure the thread goes through all guides.

**Burrs**
There may be burrs in the needle's eye, on the thread guides, needle plate or the hook. Replace the needle and try buffing the thread guides and needle plate. Buffing may alter the timing, so it's a good idea to replace a damaged hook.

**Dry Hook**
Because it dries out faster, the hook needs frequent lubrication. The hook assembly should be lubricated every 4-8 hours of machine running time.

**Needle to small**
You may have to change to a larger needle

**Too much Adhesive spray**
You should be using only Embroidery designed adhesive sprays, never over use them. Too much adhesive build up on the needle can cause them to drag and break threads. If this is happening you should examine the way you use your adhesive spray


Firefox
Embroidery Trouble Shooting Page
www.sewingandembroiderywarehouse.com/embtrb.htm
Google
zotero

### Causes:

**Improper Thread**

**Try re-threading the machine; make sure the thread goes through all guides.**

**Burrs**

**There may be burrs in the needle's eye, on the thread guides, needle plate or the hook. Replace the needle and try buffing the thread guides and needle plate. Buffing may alter the timing, so it's a good idea to replace a damaged hook.**

**Dry Hook**

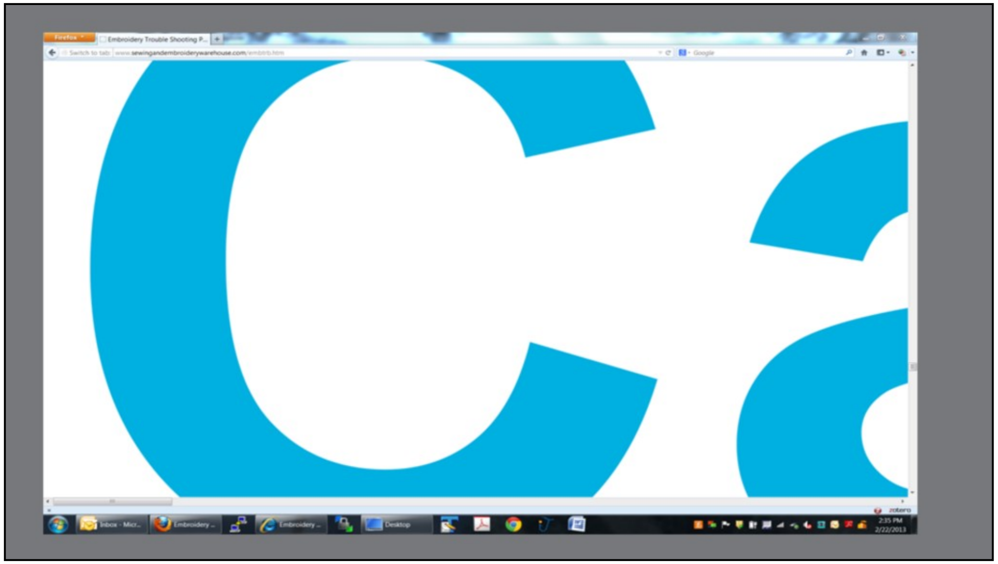
**Because it dries out faster, the hook needs frequent lubrication. The hook assembly should be lubricated every 4-8 hours of machine running time.**

**Needle to small**

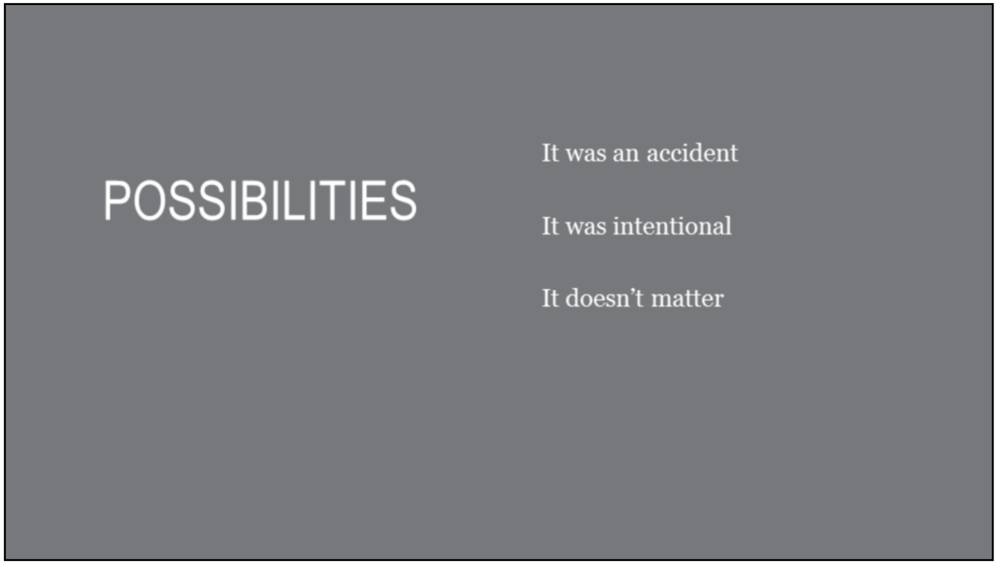
**You may have to change to a larger needle**

**Too much Adhesive spray**

**You should be using only Embroidery designed adhesive sprays, never over use them. Too much adhesive build up on the needle can cause them to drag and break threads. If this is happening you should examine the way you use**


Firefox
Embroidery Trouble Shooting P...
Switch to tab
www.sewingandembroiderywarehouse.com/embtrb.htm
Google
Ca
zotero
Inbox - Micr...
Embroidery ...
Embroidery ...
Desktop
2:35 PM
2/22/2013

*“Rendering Matters”*

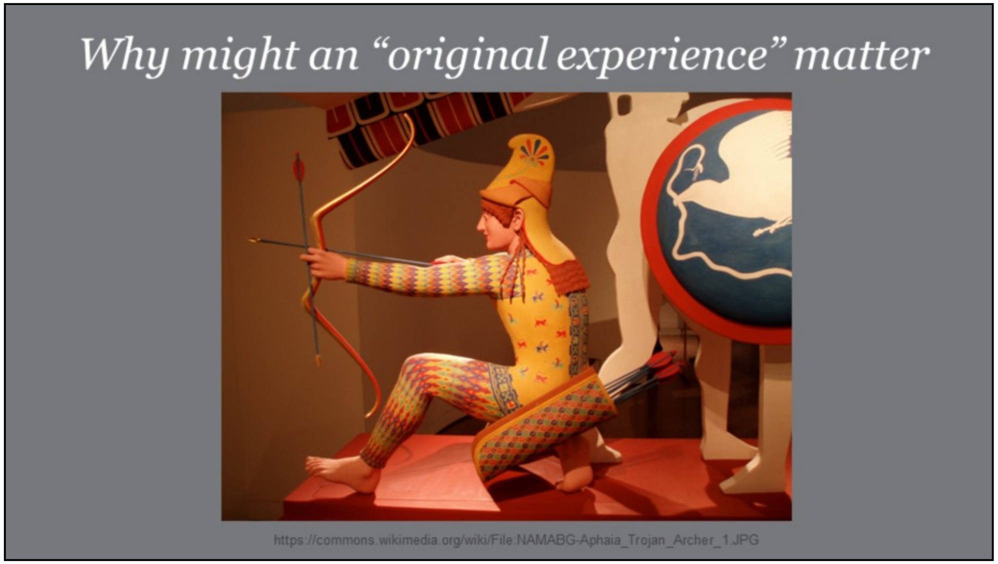


# POSSIBILITIES

It was an accident

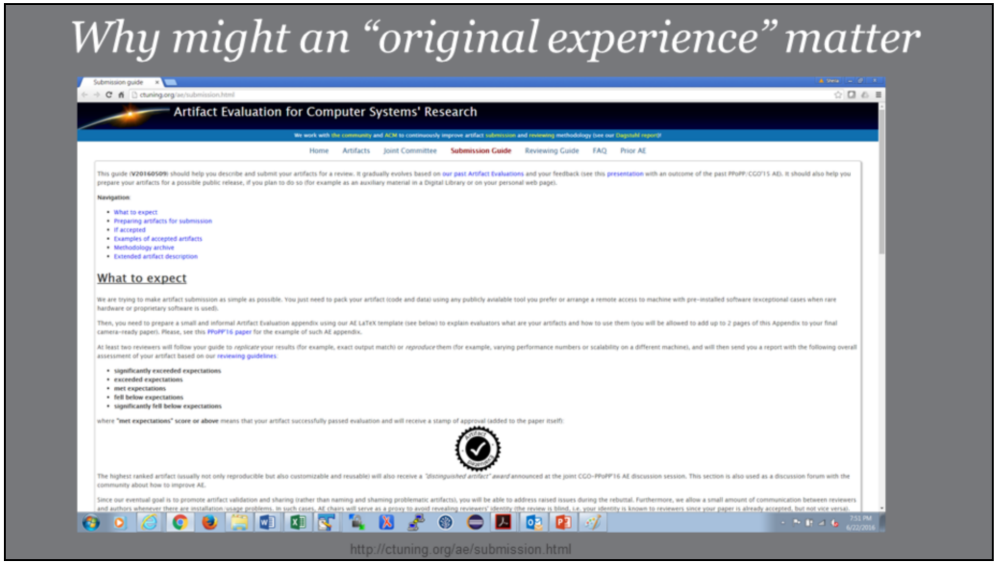
It was intentional

It doesn't matter



# Why might an "original experience" matter

https://commons.wikimedia.org/wiki/File:NAMABG-Aphaia_Trojan_Archer_1.JPG



## *Why might an "original experience" matter*

Artifact Evaluation for Computer Systems' Research

We work with the community and ACM to continuously improve artifact submission and reviewing methodology (see our Dagstuhl report)!

This guide (**V20160509**) should help you describe and submit your artifacts for a review. It gradually evolves based on our past Artifact Evaluations and your feedback (see this presentation with an outcome of the past PPoPP/CGO'15 AE). It should also help you prepare your artifacts for a possible public release, if you plan to do so (for example as an auxiliary material in a Digital Library or on your personal web page).

**Navigation:**

- What to expect
- Preparing artifacts for submission
- If accepted
- Examples of accepted artifacts
- Methodology archive
- Extended artifact description

What to expect

We are trying to make artifact submission as simple as possible. You just need to pack your artifact (code and data) using any publicly avialable tool you prefer or arrange a remote access to machine with pre-installed software (exceptional cases when rare hardware or proprietary software is used).

Then, you need to prepare a small and informal Artifact Evaluation appendix using our AE LaTeX template (see below) to explain evaluators what are your artifacts and how to use them (you will be allowed to add up to 2 pages of this Appendix to your final camera-ready paper). Please, see this PPoPP'16 paper for the example of such AE appendix.

At least two reviewers will follow your guide to *replicate* your results (for example, exact output match) or *reproduce* them (for example, varying performance numbers or scalability on a different machine), and will then send you a report with the following overall assessment of your artifact based on our reviewing guidelines:

- **significantly exceeded expectations**
- **exceeded expectations**
- **met expectations**
- **fell below expectations**
- **significantly fell below expectations**

where **"met expectations" score or above** means that your artifact successfully passed evaluation and will receive a stamp of approval (added to the paper itself):

The highest ranked artifact (usually not only reproducible but also customizable and reusable) will also receive a *"distinguished artifact" award* announced at the joint CGO-PPoPP'16 AE discussion session. This section is also used as a discussion forum with the community about how to improve AE.

Since our eventual goal is to promote artifact validation and sharing (rather than naming and shaming problematic artifacts), you will be able to address raised issues during the rebuttal. Furthermore, we allow a small amount of communication between reviewers and authors whenever there are installation/usage problems. In such cases, AE chairs will serve as a proxy to avoid revealing reviewers' identity (the review is blind, i.e. your identity is known to reviewers since your paper is already accepted, but not vice versa).

http://ctuning.org/ae/submission.html

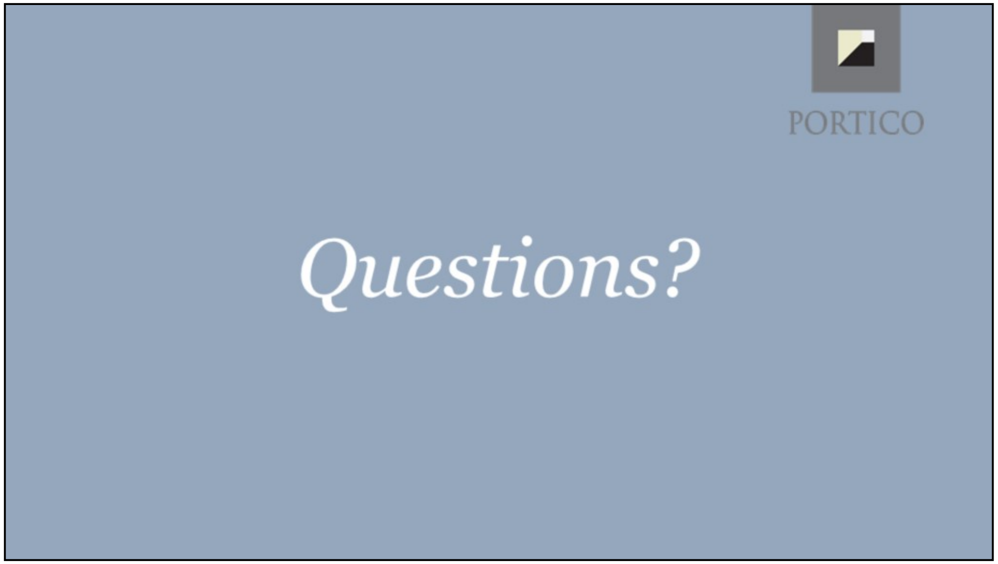
**Data, Software, & Reproducibility**

This is a joint Project of the ACM Digital Library and the ACM Technology Committees

*Meanings matter*

The meaning and value of these data derive not only from the raw content, but from the ways people interact with the technologies that create them

Sara Day Thompson – preserving transactional data



PORTICO

*Questions?*