

The Craft of XML

Text Encoding in historical and humanistic context

Wendell Piez
JADH 2015
University of Kyoto
Kyoto, Japan
September 2 2015



An *Ukiyo-e* woodblock print depicting a woodblock printing shop.
(Is the scene realistic, or fanciful?)
Utagawa Kunisada (1786–1865)

What is “craft”?

工 こう

芸 げい

Haiku Demo - Chromium

M Ka × 兵 × Se × H JA! × Q we × Hc × Ja! × Ta × Ha ×

file:///home/wendell/Documents/JADH2015/HaiKuML/xml/ba:

Basho Matsuo

Even in Kyoto
To be longing for Kyoto:
A cuckoo cuckoos.

時 ほどと 鳥 ぎす 京 きょう なつかし 京 きょう にても や

Themes: Summer; Hotogisu

Vocabulary

なつかし
Nostalgia; missing someone or something.

Remarks

Interestingly, hotogisu is a "time bird". This may give the verse an additional (if laconic) poignancy.

Links:

<http://carlsensei.com/classical/index.php/text/view/54>

An audio recording of a hotogisu (Youtube) [https://www.youtube.com/watch?v=2F_KfMB5ICU]

*And what is XML?
And what does this mean
for the humanities?*

Kanji dictionary

Consulting the Digital Dictionary of Buddhism

京

Metropolis, capital. [首都 帝都] (Shogakukan)

reading: ケイ

A (high, large) hill. [丘] (Shogakukan)

reading: キョウ

Large, great, flourishing, tall. [大 高] (Shogakukan)

reading: kei

A storehouse; a granary. [倉] (Shogakukan)

reading: kyō

A numerical power of ten. Anciently, ten times one 卍. In

reading: kin

In the tenth century, 一 (ichi).

reading: miyako

reading: takai

Craft vs Industry



Raku bowl (Kyoto, 18th-19th Centuries)
Freer Gallery of Art, Washington DC
(Picture from Wikimedia Commons)



Craft

Sensitive to history, materials, purpose
Seeks distinctive virtue in each production
Meaning is in materiality

Industry

All about regularity, scalability
Keep the costs down!
No surprises
One is as good as another

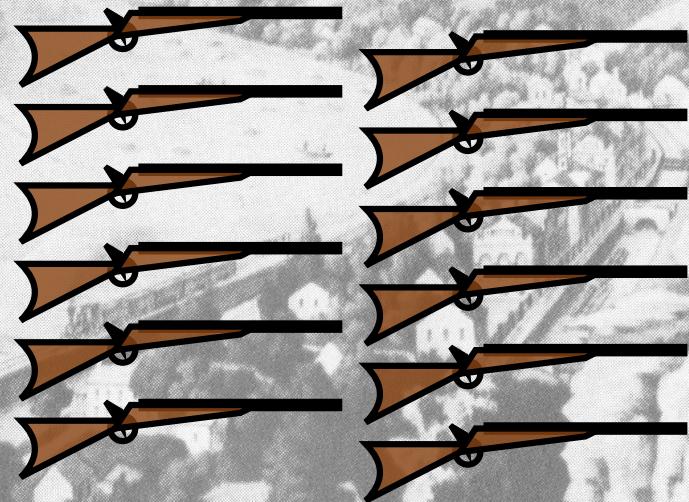
<https://pixabay.com/en/history-pottery-shells-blue-64971/>

Craft *versus* / and Automation

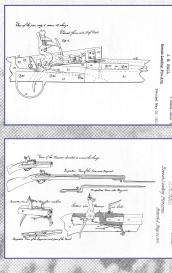
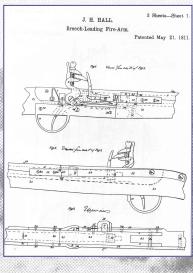
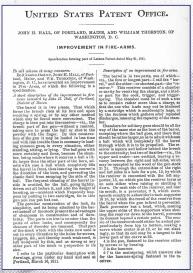
	Purpose	
	<i>Purpose of maker in service to recipient</i>	   
	<i>Sensitivity and celebration</i>	   
	<i>Consciousness and dialog with history and tradition</i>	   
	<i>Perfection in imperfection (cf. wabi-sabi)</i>	   
	<i>Acceptance of time and transition / temporariness</i>	   

Automation Takes Flight

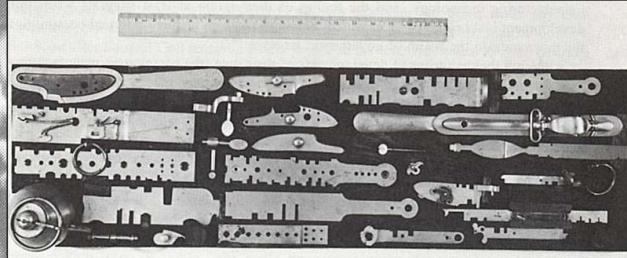
Harper's Ferry, Virginia, 1818



Maine gunsmith Captain John Hall contracts with the US Army to produce rifles with interchangeable parts. This is only possible by automating production and controlling fabrication by machine.



The method is *measurement* with reference to an *abstract model*



Abstract Specifications

J. H. HALL.
Breech-Loading Fire-Arm.

Patented May 21, 1811.



All inputs and processes are codified, normalized and controlled.

Inputs include all necessary resources (time, materials, labor).

Outputs are described and specified before they are made.

This principle can be applied to *any* kind of production

(not just gunsmithing).

J. H. HALL.
Breech-Loading Fire-Arm.

3 Sheets—Sheet 2
Patented May 21, 1811

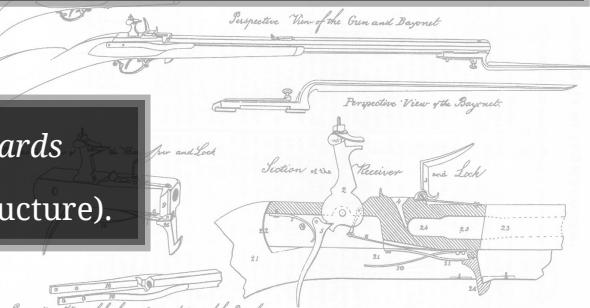
J. H. HALL.
Breech-Loading Fire-Arm.

J. H. HALL.
Breech-Loading Fire-Arm.

Sheets 3, 3 Sheets;
Patented May 21, 1811

Bicycles, sewing machines, books, printing presses ...

Formalizing specifications also permits *standards* and *commodity markets* (on a shared infrastructure).



Fast Forward >>> to 1970s-80s

The digital information processor (aka “computer”) is the culmination of automation technologies: the *universal machine*.

The problem:

What if your information is
rare, expensive, valuable, ?
(And your computer is dead in 5 years?)

The solution:

Non-proprietary information technologies: *open standards*
Providing a basis for

- Platform independence
- One data set, many applications
- Sharing of knowledge and expertise

SGML (Standard Generalized Markup Language) released in 1986.



Almost 40 years ago, I worked
on a computer like this.
Regrettably, nothing survives.

Principles of Generalized Markup

XML (TEI) example at

```
<body><a href="http://www.piez.org/wendell/projects/buechlein/fechner-edited.xml">  
<pb n="1"/>  
<head type="main">LIFE AFTER DEATH</head>  
<div>  
    <head>CHAPTER I</head>  
    <p>MAN lives upon the earth not once, but  
        of life is a continuous sleep; the second  
        sleeping and waking; the third is an eter  
        >in the first stage man lives alone in d  
        hear and among others, but detached and i  
        the third his life is merged with that of  
        >the body is developed  
        its equipment for the second; in the sec  
        its seedbed and realizes its powers for t  
        >the spark which lies in  
        already here through perception, faith, f  
        Genius demonstrates the world beyond man  
        stage as clear as day, though to us obscu  
        <p>The passing from the first to the secon  
        second to the third is called death.</p>  
        <p>The way upon which we pass from the sec  
        not<pb n="3"/> darker than that by which  
        first. The one leads to the outer, the ot  
        the world.</p>
```

Establish a base line character set (e.g., Unicode)

Agree on a markup syntax (e.g., XML)

Present data (information) as mix of *text* ("content") and *markup*

Differentiate between data for process(es) and end user(s)

Typically, text is for users, and markup is for processes

(This line can be fuzzy.)

Deploy system in *layers*:

Processes can work with markup and/or data as appropriate

Information can be differentiated for querying

Markup and content can be tested and validated separately

(So roles of people dealing with each can also be defined.)

Out-of-line processing (e.g. stylesheets) can be applied without modifying sources

Use markup to *describe* data

Markup semantics can be application-independent

The Layered Architecture of XML

XML parser reads markup from file or bitstream, and builds a model.

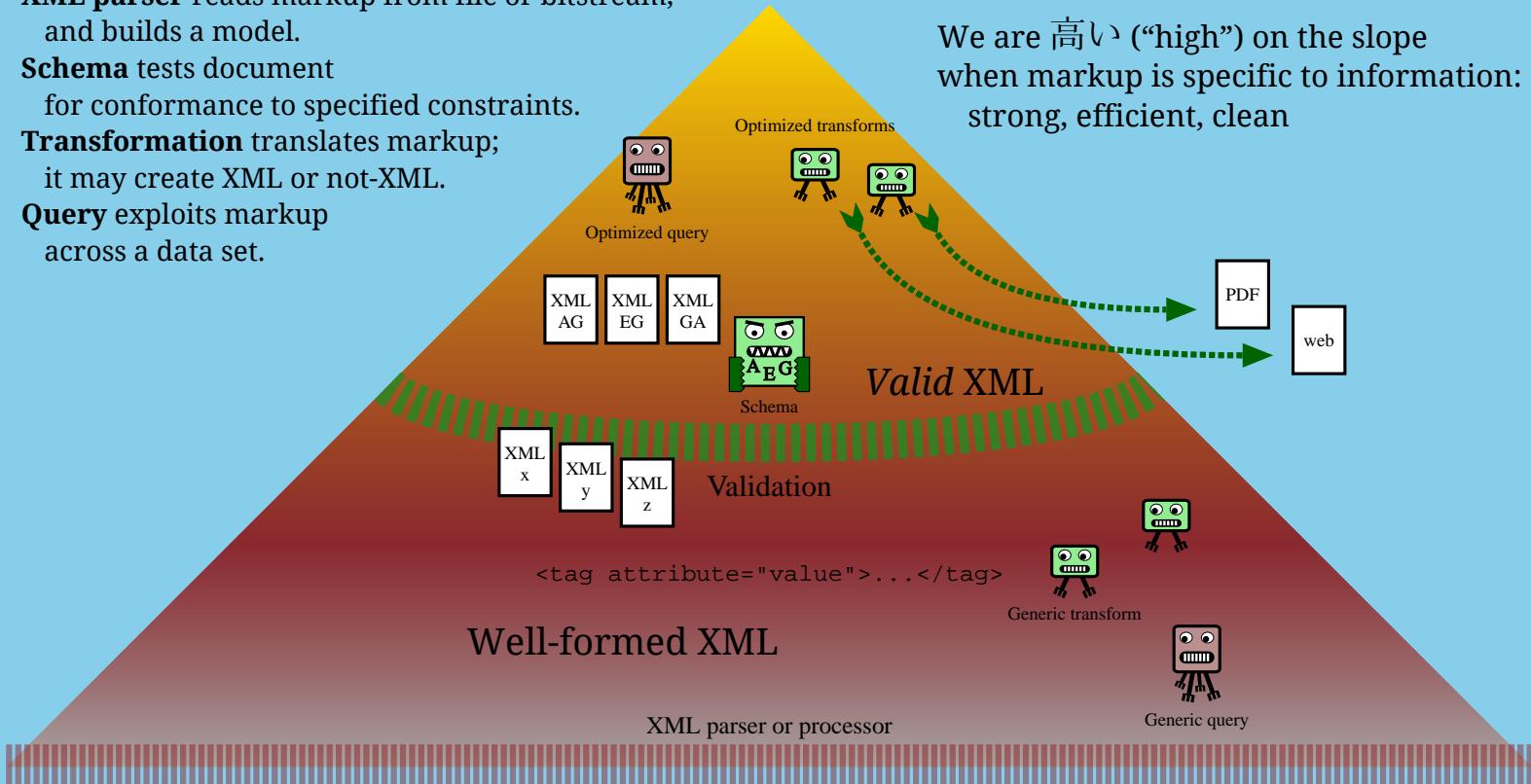
Schema tests document for conformance to specified constraints.

Transformation translates markup; it may create XML or not-XML.

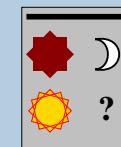
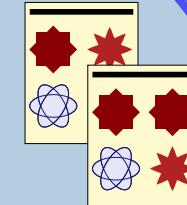
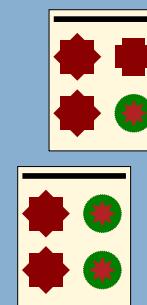
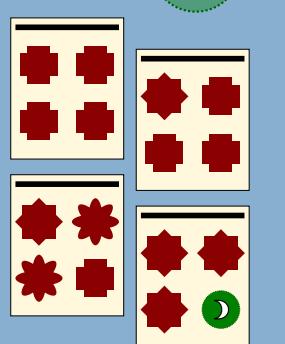
Query exploits markup across a data set.

To go “up hill” is difficult
To go “down hill” is easy;

We are 高い (“high”) on the slope when markup is specific to information: strong, efficient, clean



A schema defines a boundary line
between known and unknown



This makes it possible to develop processes
Before we have seen all the data

But ... which XML do I use? ...

(... for example ...)

TEI

Text Encoding Initiative

Produced by an academic consortium (tei-c.org)
Proposes tagging for digital humanities projects
Large, complex, more than anyone needs
(But what you need might be in it!)

JATS

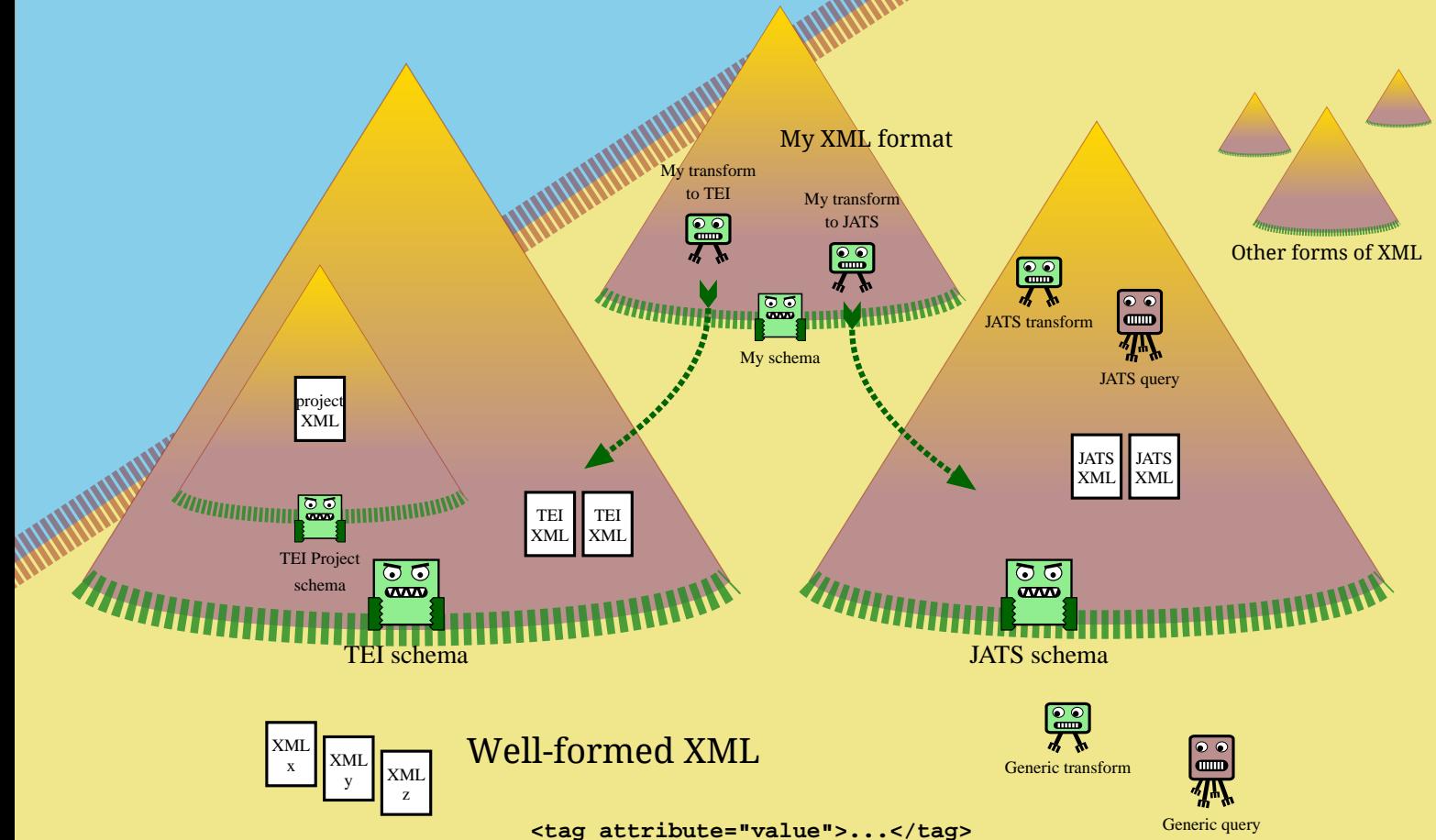
Journal Article Tag Suite

Originally produced at NIH/NLM
(US National Library of Medicine
at the National Institutes of Health)
Codifies common practice in journal publishing
Now standardized at NISO
(National Information Standards Organization, USA)
Specifically for journal publishing
Also now book publication! (BITS)
More common in commercial publishing
Especially scientific/technical/medical publishing
Easier to use than TEI (half the complexity)
Conference in Tokyo next month!
JATS-Con Asia (see <http://xspa.jp/>)

Or ... something else?
(EAD, METS/MODS, Docbook, DITA, etc. etc. ...?)

Or ... *design your own XML?*

Varieties of XML



Craft After All?

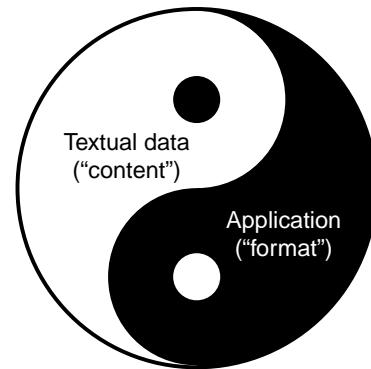
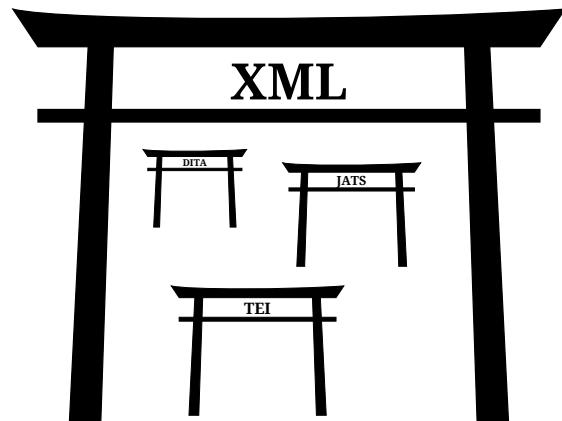
XML text encoding technologies in the service of applications in the humanities

Avoid proprietary entanglements

Data and application are not separate, but married

The machine (the medium) matters!

A standard is not an end point, but a gateway



The Craft of XML by Wendell Piez

JADH 2015, University of Kyoto

Kyoto, Japan, September 3 2015