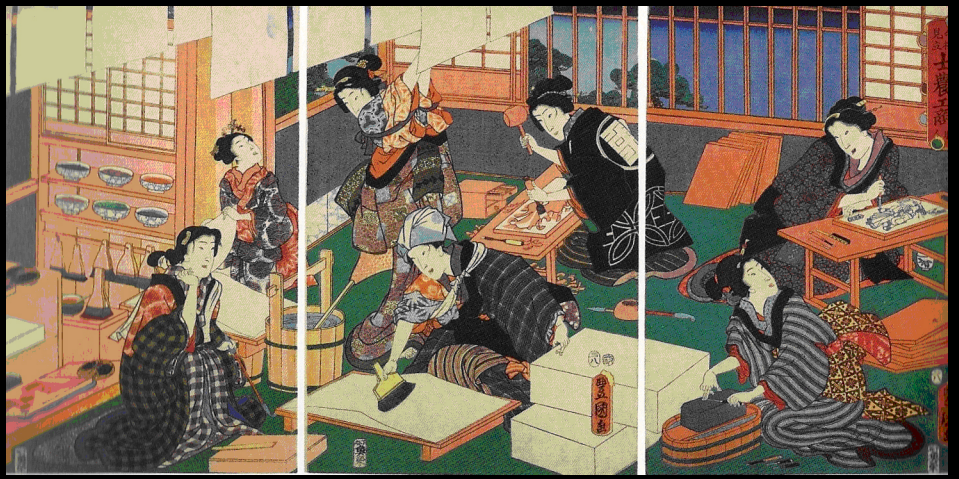
The Craft of XML

Text Encoding
in historical and humanistic context

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

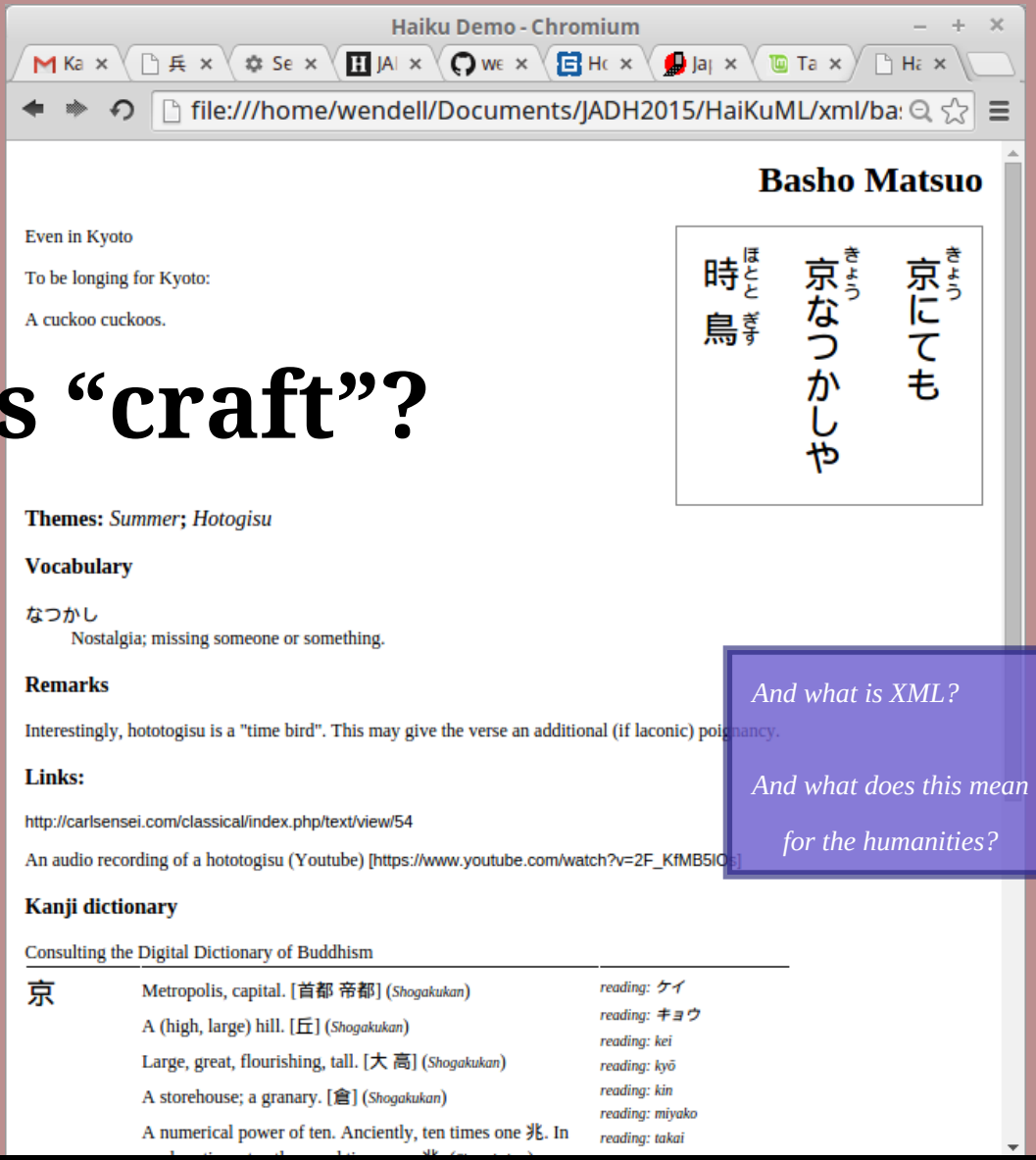


An *Ukiyo-e* woodblock print depicting a woodblock printing shop.
(Is the scene realistic, or fanciful?)
Utagawa Kunisada (1786–1865)

What is “craft”?

工 こう

芸 げい



And what is XML?

And what does this mean for the humanities?

Craft vs Industry

Craft

Sensitive to history, materials, purpose

Seeks distinctive virtue in each production

Meaning is in materiality

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

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 of maker
in service to recipient

Purpose

No purpose or any purpose
(E.g., sell a bunch of stuff)

Sensitivity
and celebration

Materials

Material is treated as input
(with time and labor)

Consciousness and dialog
with history and tradition

History

No history, past or future,
Only sequence of operations

Perfection in imperfection
(cf. wabi-sabi*)*

Perfection

Optimization
making choices among tradeoffs

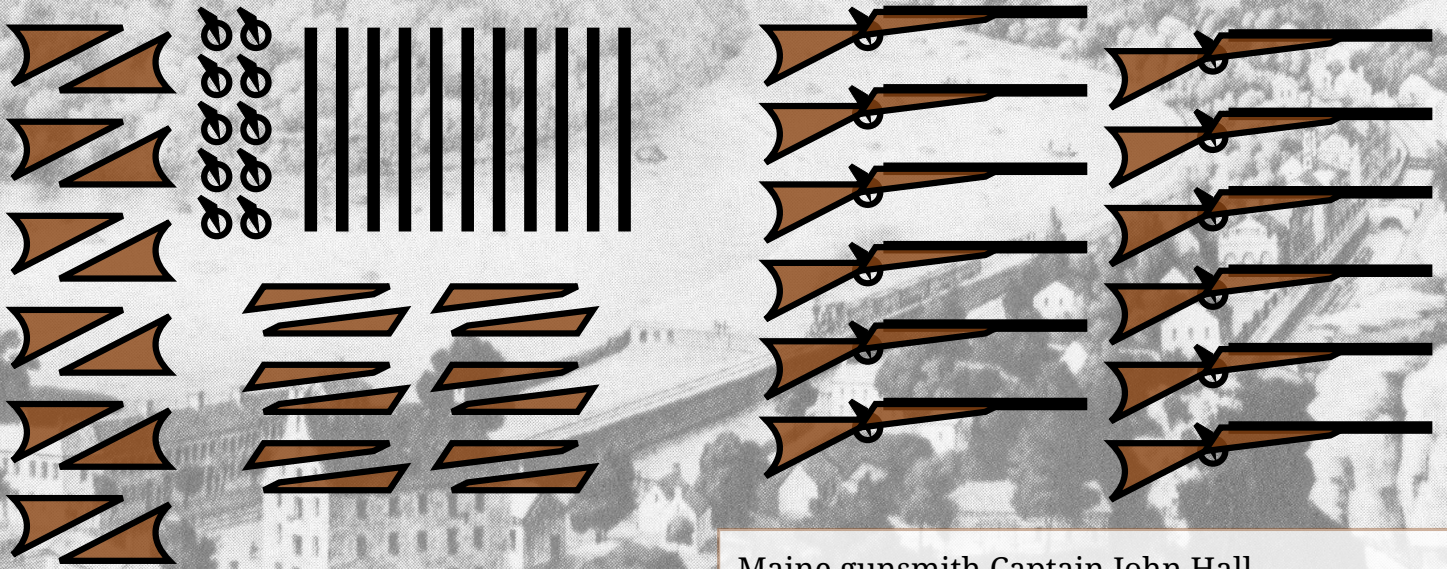
Acceptance of time
and transition / temporariness

Time

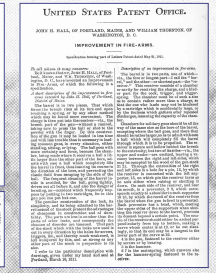
No more time
only duration (a resource)

Automation Takes Flight

Harper's Ferry, Virginia, 1818



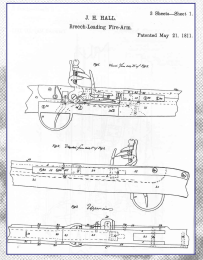
Maine gunsmith Captain John Hall
contracts with the US Army
to produce rifles with interchangeable parts.
This is only possible by automating production
and controlling fabrication by machine.



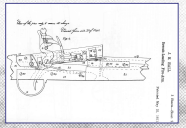
UNITED STATES PATENT OFFICE.

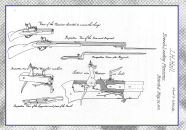
JOHN H. HALL, OF PORTLAND, MAINE, AND WILLIAM THORNTON, OF WASHINGTON, D. C.

IMPROVEMENT IN FIRE-ARMS.

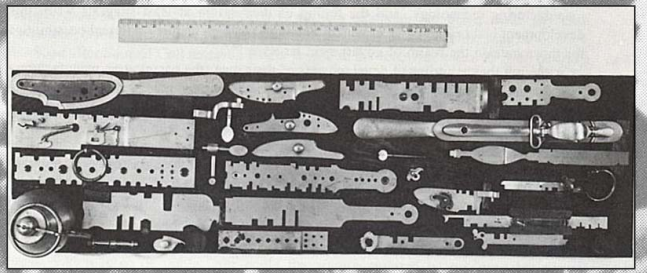


J. H. HALL.
Breech-Loading Fire-Arm.
No. 1. 4 Sheets—Sheet 1.
Patented May 21, 1811.





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



Abstract Specifications

All inputs and processes are codified, normalized and controlled.

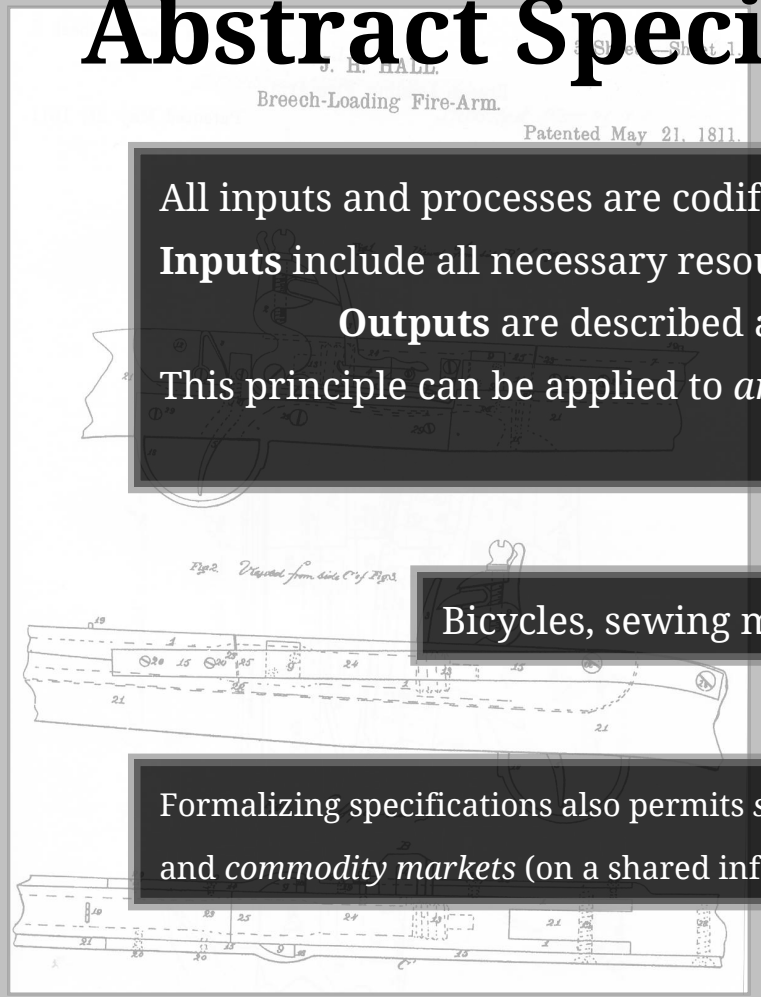
Inputs include all necessary resources (time, materials, labor).

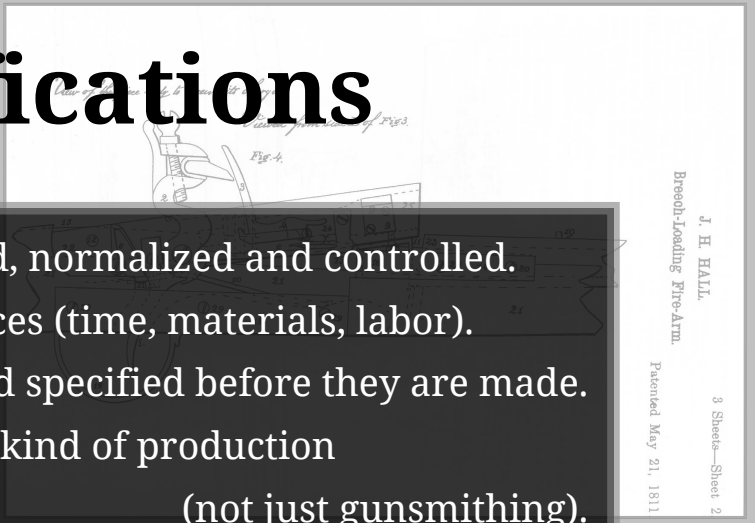
Outputs are described and specified before they are made.

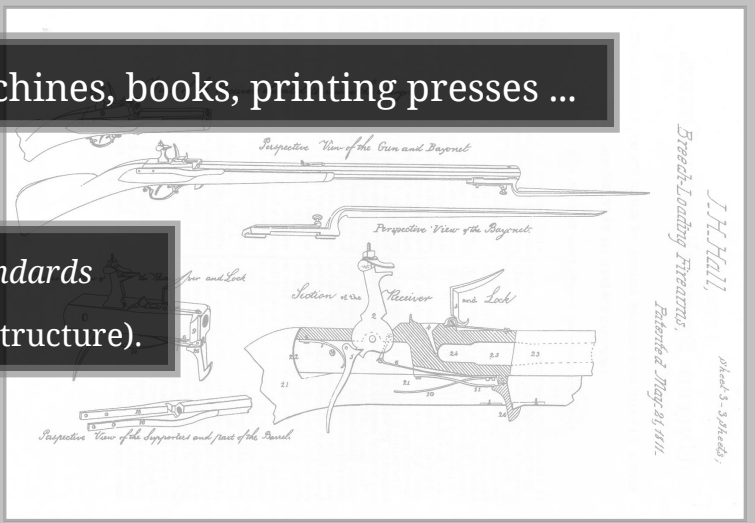
This principle can be applied to *any* kind of production (not just gunsmithing).

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*.

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

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.

Principles of Generalized Markup

XML (TEI) example at http://www.piez.org/wendell/projects/buechlein/fechner-edited.xml

```
<body>
 <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
   <p>In the first stage man lives alone in d
   near and among others, but detached and i
   the third his life is merged with that of
   <pb n="2"/> Spirit, and he discern
   <pb n="2"/> stage the body is develope
   its equipment for the second; in the seco
   its seedbud and realizes its powers for t
   developed the divine 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

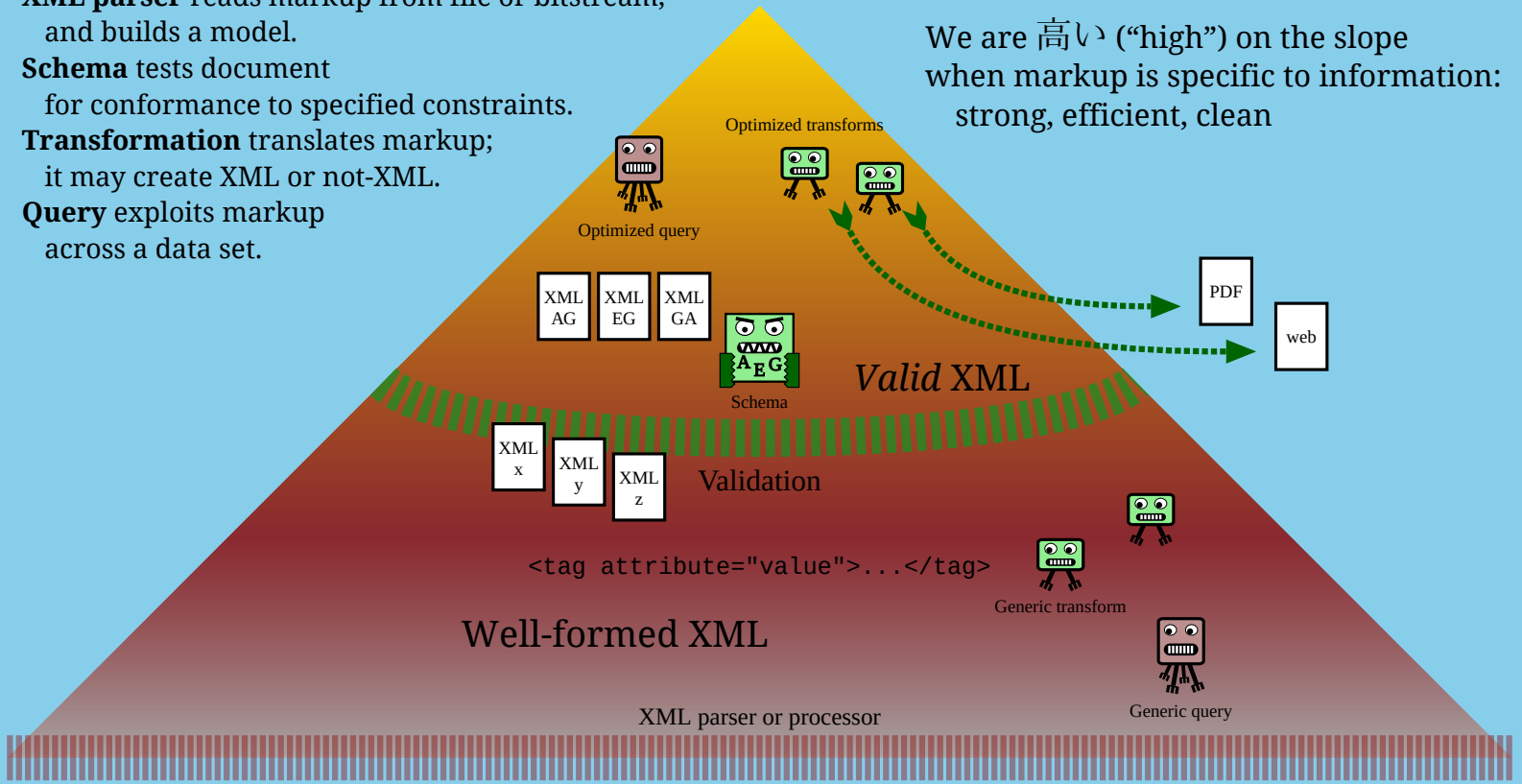
- Markup semantics can be application-independent

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.

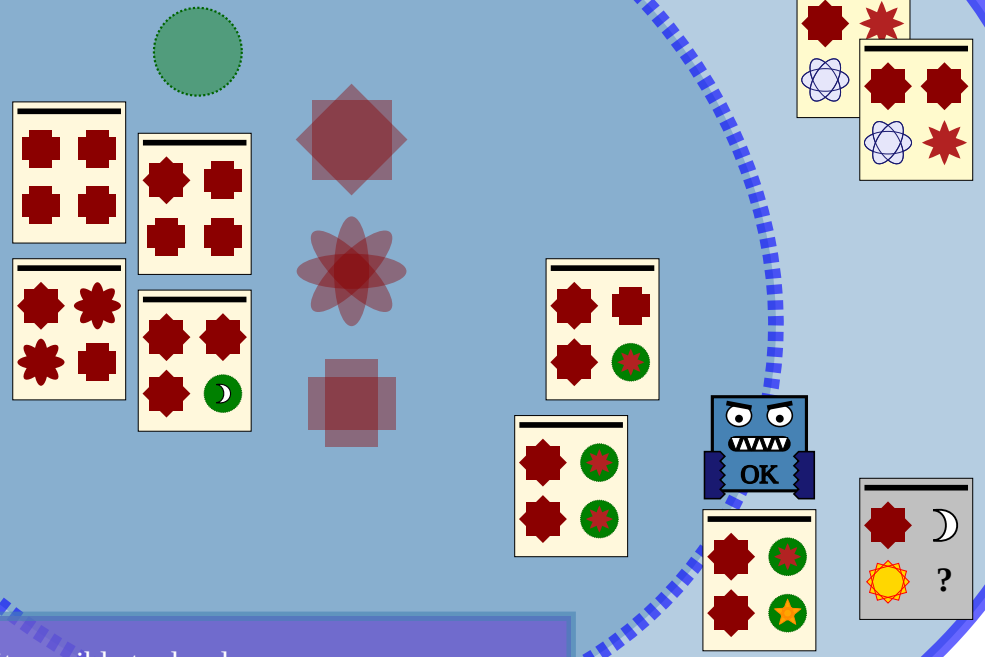
To go “up hill” is difficult
To go “down hill” is easy;

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 **E**ncoding **I**nitiative

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 **A**rticle **T**ag **S**uite

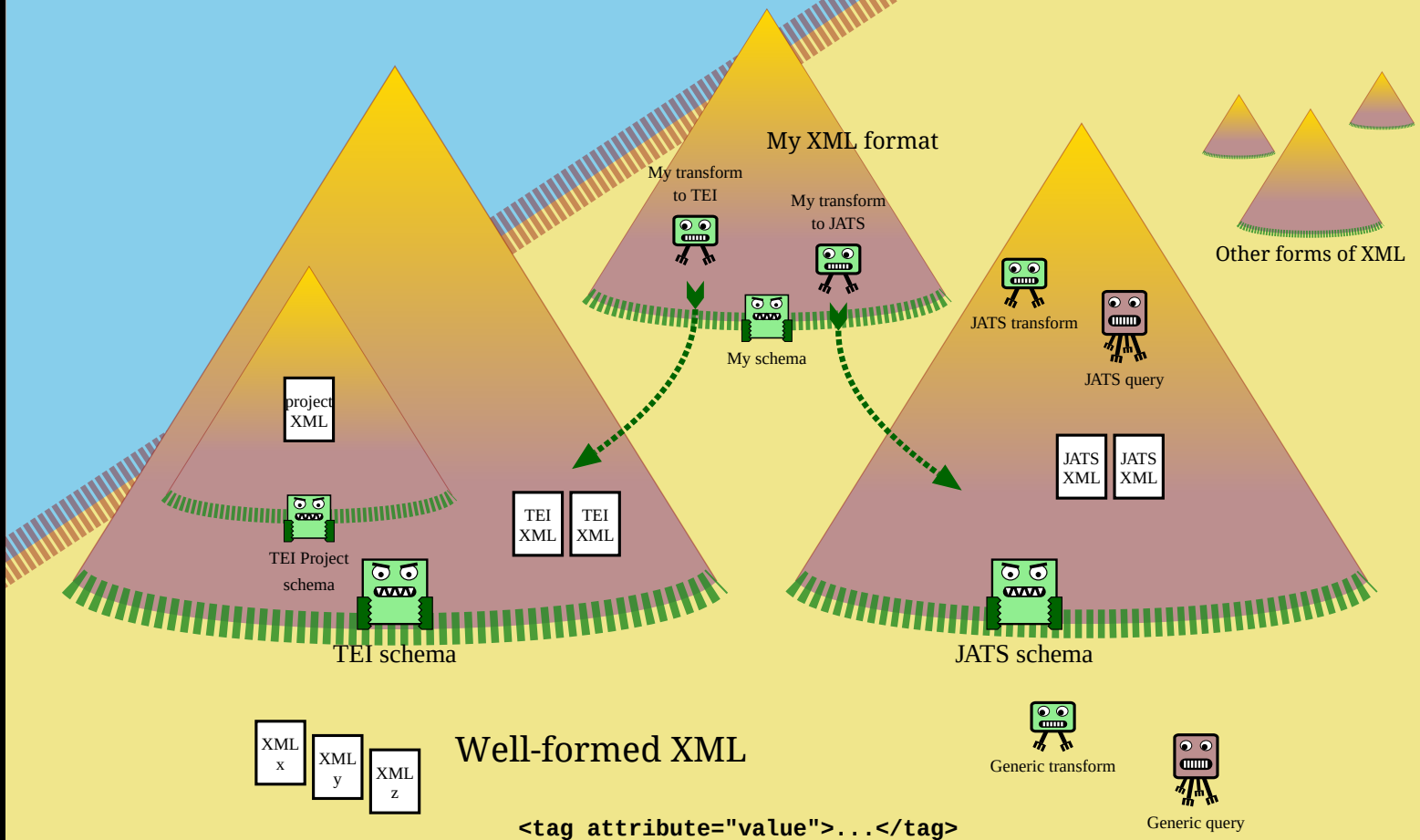
Originally produced at NIH/NLM
(US National Library of Medicine
at the National Institutes of Health)
Codifies common practice in journal publishing
Now standardized at NISO
(National Information Standards Organization, USA)
Specifically for journal publishing
Also now book publication! (BITS)
More common 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?

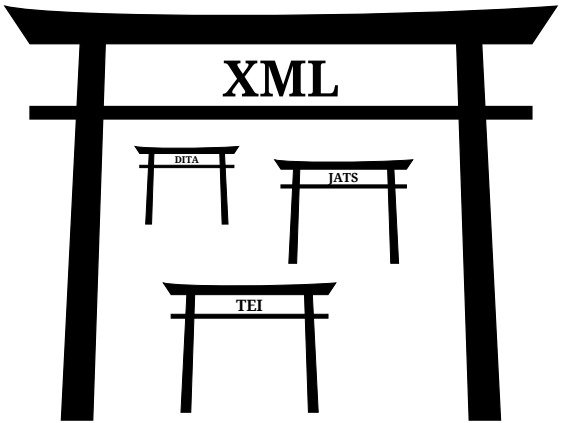
XML text encoding technologies in the service of applications in the humanities

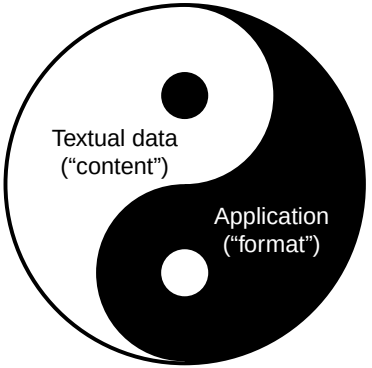
Avoid proprietary entanglements

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 2 2015