

XML と JATS

2022/10/12

東京大学大学院情報学環高等客員研究員

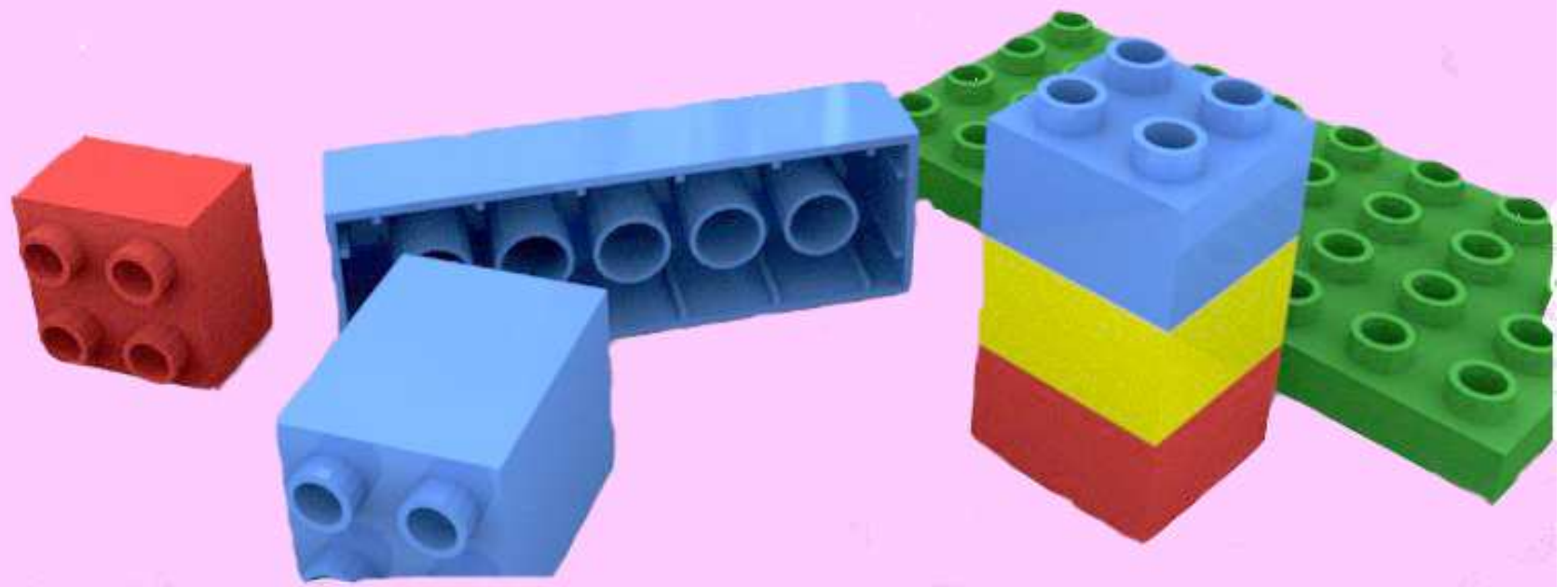
学術情報 XML 推進協議会会長

時実 象一

tokizane@pc.highway.ne.jp

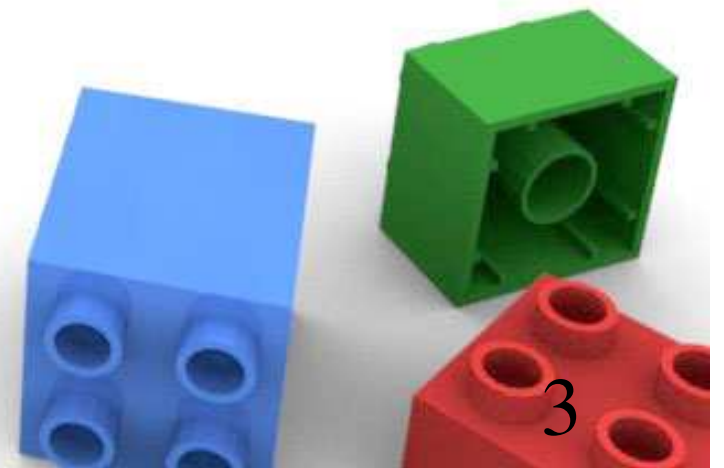


XML (eXtensible Markup Language)



Markup

- テキストのスタイルづけ (組版指定)
 - TR36b/c Times Roman 36 point bold, centered
(<http://www.webreference.com/authoring/design/artsci/chap1/1/2.html>)
 - 本文 8 ポ 2 段組、1 行 25 字 x 45 行、行間 5 号 2 分、カンマとピリ



本文 12Q. 2段組. 24字×44行. 20H, カンマビリ

結果^{および}考察 ↑ 2Lビ4 左中央 (Table)

Table 1 調査対象の平均身長^{および}及体重を~~表~~1に示した。
厚生省国民栄養調査報告^{および}と比較すると身長は男女共、各年代共にほぼ近似してゐるが、
体重は調査対象側の^{ほう}が軽い傾向にある。こ

|| Q (〇本論文の要旨は第68次日学協会年会論文(久留米)におい^てて表した。) 2段並行止X

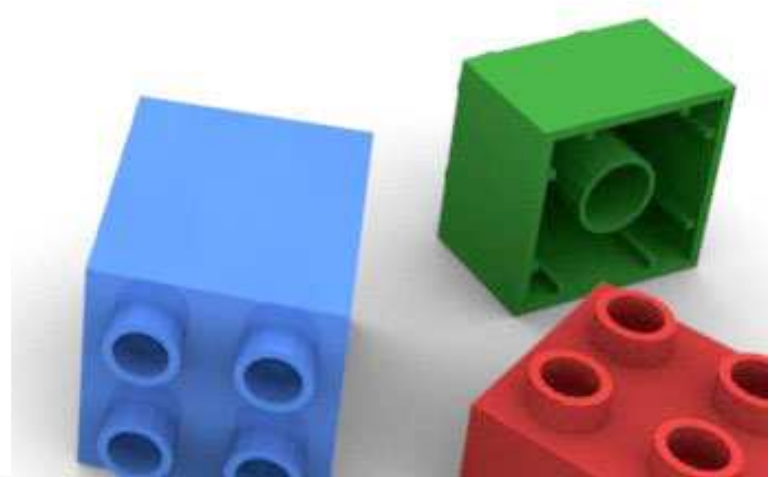
↓ || Q 文〇〇〇献 ↑ 2Lビ4 全段左右中央

1 参考文献

14 H ↓

- 1) Aimi's (1952) Studies on the weight and size of internal organs of Normal Japanese. Acta Pathologica Japōnica 20:173-1200. ^{289-2 (以F刊)}
- 2) 平沢^{おと} 櫻^{おと} 園^{おと} 下^{おと} 達^{おと} 雄 (1969). 解剖学 2巻第10版. 東京, 金原出版, p. 6.
- 3) 伊藤^{おと} 隆 (1983). 解剖学講義. 東京, 南山堂, p. 279.
- 4) 金子^{おと} 丑^{おと} 之^{おと} 郎 (1982). 日本人の解剖学.

目版例



Runoff (清書)

- 組版言語

.line length n

.indent n

.single space

.begin page

.center

1 行の長さを n に

インデント文字数を n に

行間を 1 行に

新しいページを始める

中央揃え

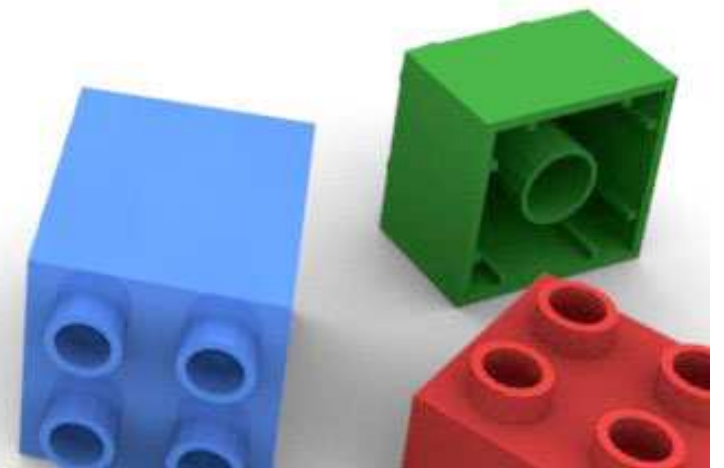
<http://mit.edu/Saltzer/www/publications/CC-244.html>



Scribe

- Brian Reid (Carnegie Mellon University)
(1980)

```
@Heading(The Beginning)
@Begin(Quotation)
    Let's start at the very beginning, a very good place to start
@End(Quotation)
```



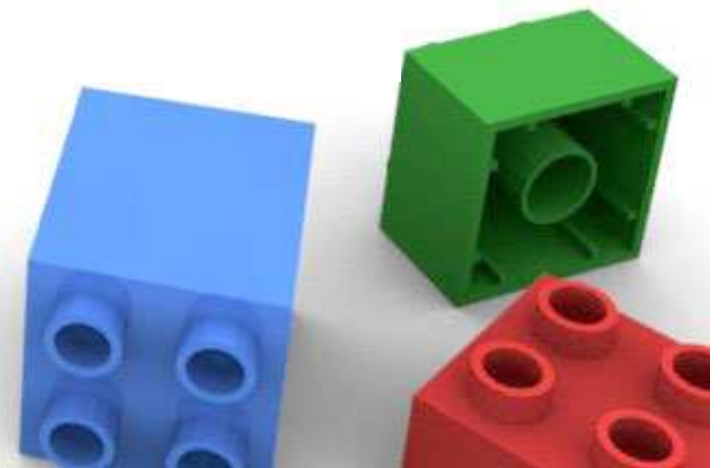
LaTeX

```
\begin{align}
E &= mc^2 \\
m &= \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}} \\
\end{align}
```



$$E = mc^2 \tag{1}$$

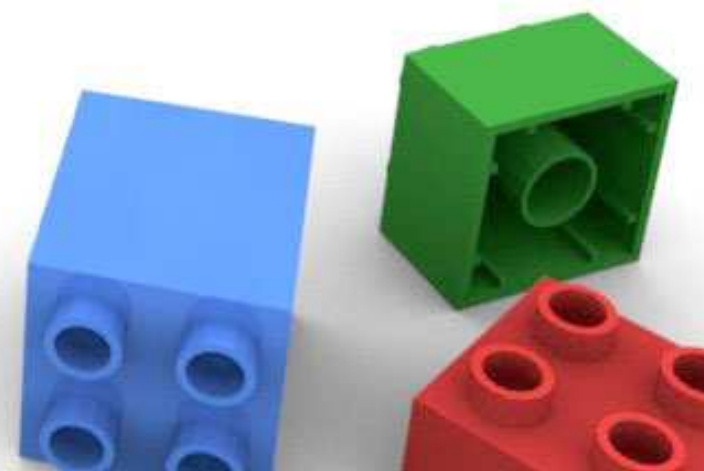
$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2}$$



Postscript

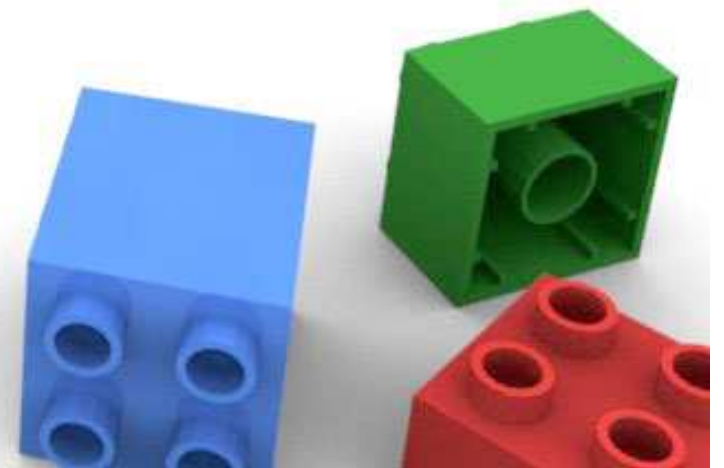
```
/Times-Roman findfont % Get the basic font
20 scalefont % Scale the font to 20 points
setfont % Make it the current font

newpath % Start a new path
72 72 moveto % Lower left corner of text at (72, 72)
(Hello, world!) show % Typeset "Hello, world!"
```



SGML

- Standard Generalized Markup Language
 - 元は IBM's Generalized Markup Language
 - Charles Goldfarb, ほか (1960 年代)



SGML

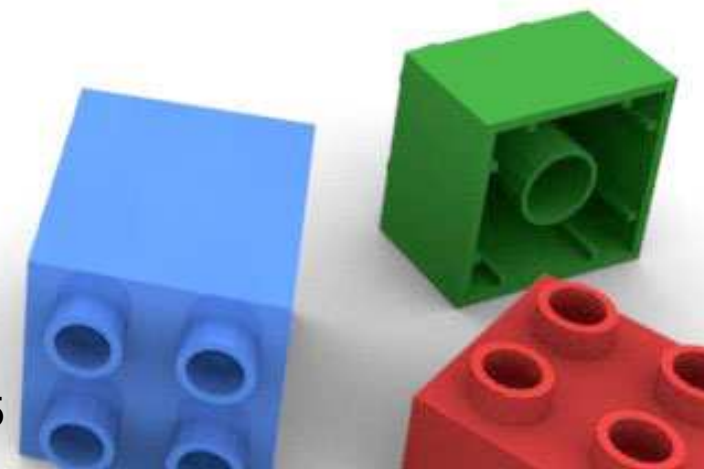
```
<point>
<title>What is the value of adding explicit structure descriptors
to text?</title>
<text>The rewards of adding explicit structure descriptors to text
are many, including:
<list>
<item>Stimulating reflection on the logical structure of what one
is saying, leading to:
<list>
<item>Better style: Making what one is trying to say clearer
to the persons who will read and try to understand it</item>
<item>Clearer thinking: Forcing oneself to become clearer
what one is trying to say</item>
<item>Discovery: Discovering things about both the content and
the form of expression which one had not previously thought
of, through the process of articulating what one thinks
one wants to say</item>
</list></item>
```

- <http://www.users.cloud9.net/~bradmcc/WhatIsSGML.html>



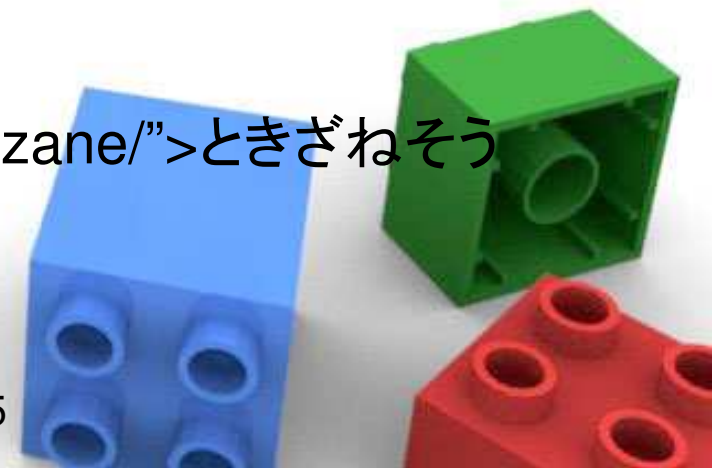
SGML, HTML, XML

- SGML (Standard Generalized Markup Language)
 - 元は IBM's Generalized Markup
 - 政府文書の作成などで普及
 - 学術雑誌の編集も
- HTML (HyperText Markup Language)
 - SGML の Web 版とみなされる
 - DTD が不要



HTML とは

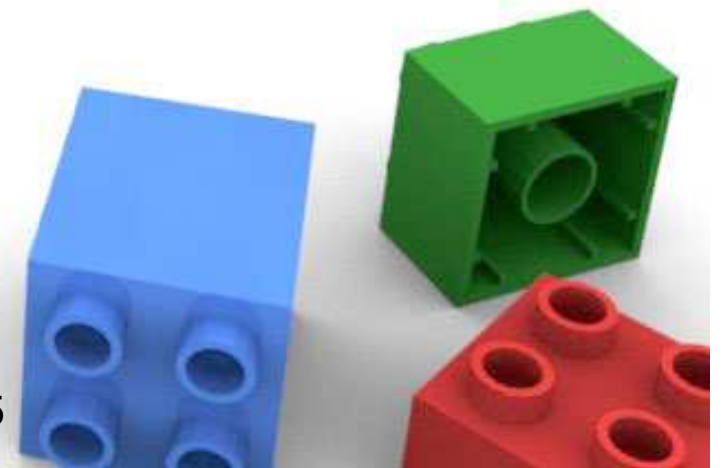
- Web ページを記述する言語
 - レイアウトを指定できる
 - 画像などを貼ることができる
 - リンクを張ることができる
- 基本的なスタイル
 - <開始タグ><終了タグ>からなる
 - ときざね
 - ときざね
 - リンク
 - ときざねそう
いち



XML とは

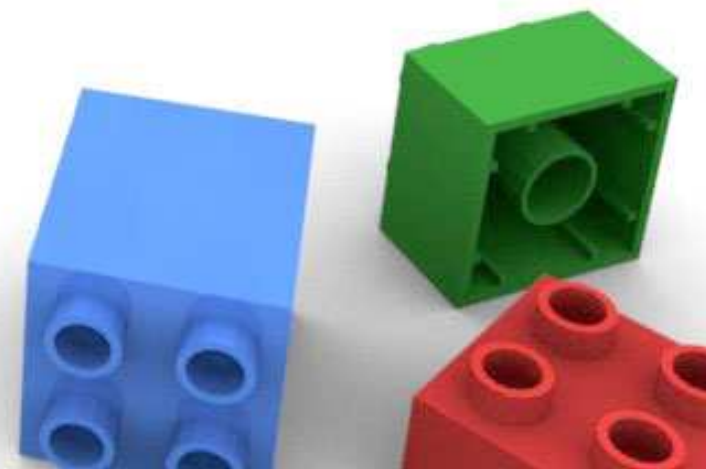
- XML (eXtensible Markup Language)
 - データの構造を表現する言語
 - 表示のためには CSS (Cascading StyleSheet) などを使用

<http://www6.airnet.ne.jp/manyo/xml/>



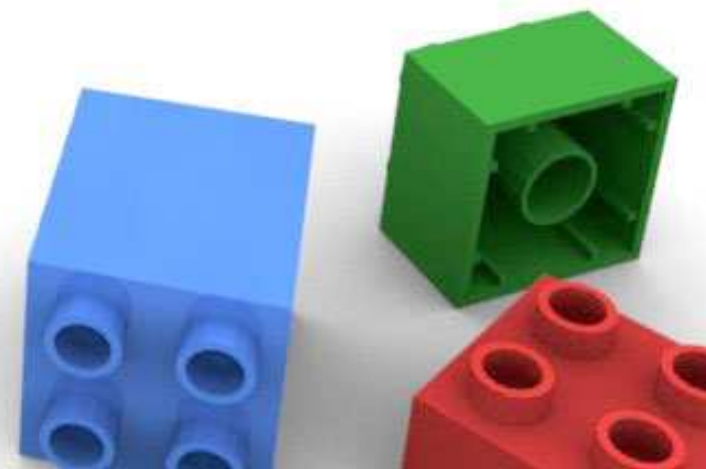
XHTML

- eXtensible HyperText Markup Language
- SGML の Web 版
- well-formed でなくてはならない
 - <tag> には </tag> が対をなす
 - case-sensitive
 - HTML の略記法は使えない
- XML パーサーでチェックできる



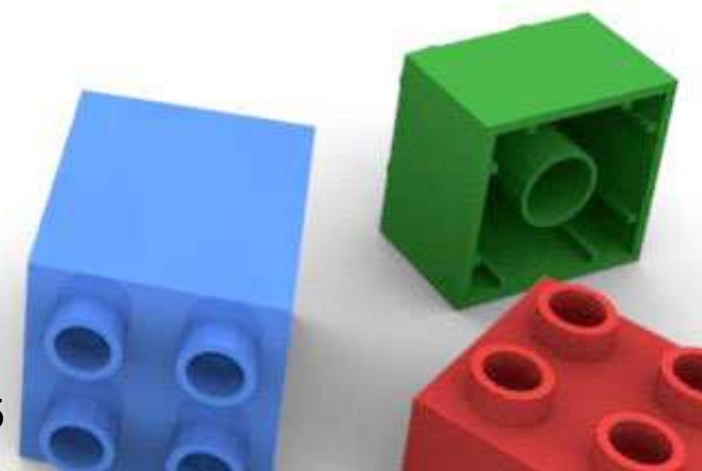
電子書籍

- ePUB
 - zip ファイル
 - META-INF (Dublin core)
 - OPS (コンテンツは XHTML または DTBook)
 - PDF



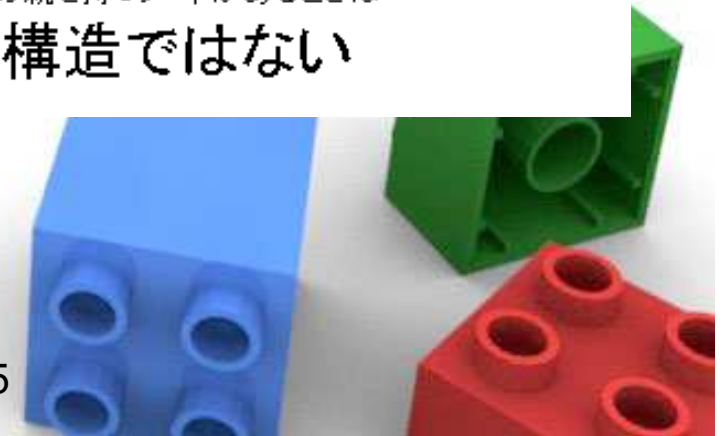
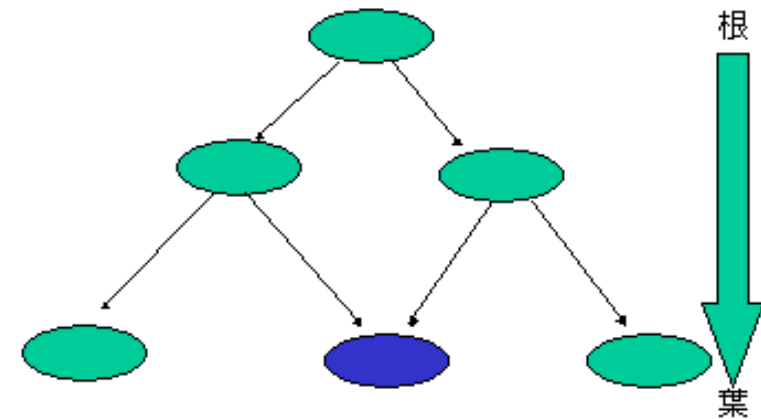
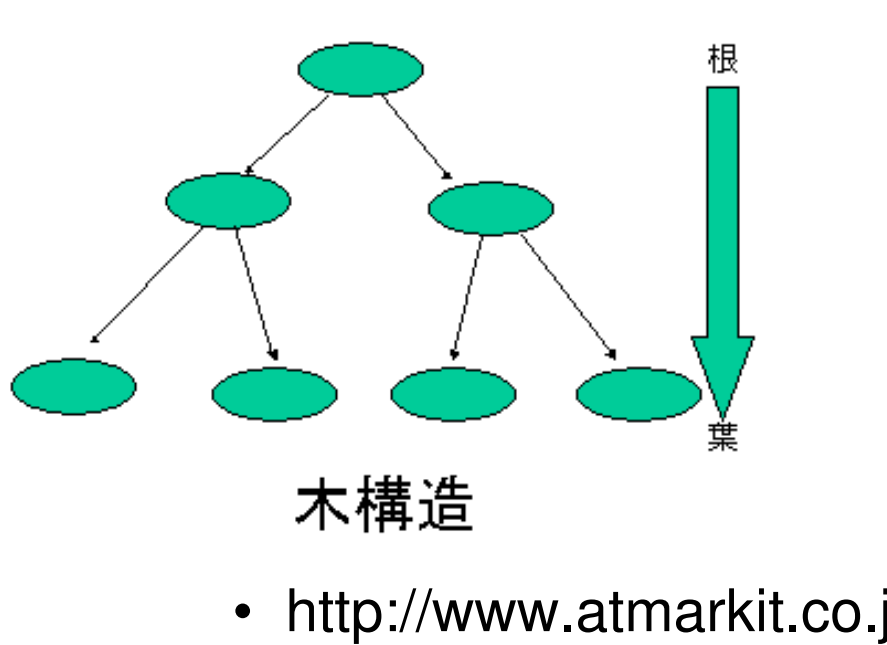
XML

- eXtensible Markup Language
 - 現在電子文書の中心
 - データ交換
 - メタデータ記述 (RDF)
 - 電子ジャーナル
 - 電子書籍 (XHTML, EPUB)
 - Microsoft Office
 - InDesign



構造

- XML の構造は木構造 (tree structure) である



要素 (Element)

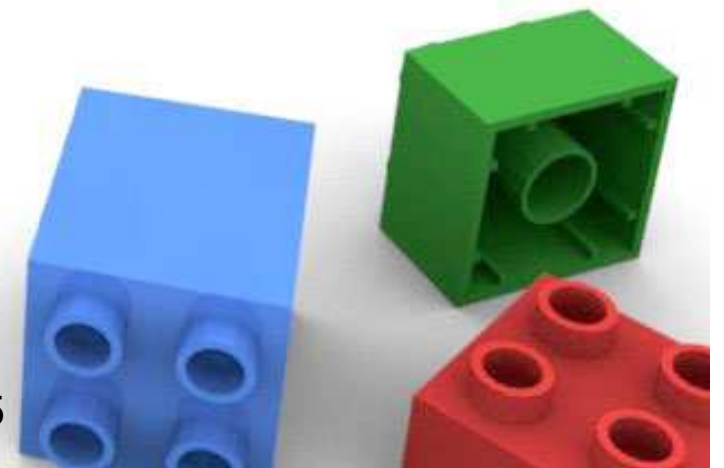
- XML の情報単位
- 「物」を表す
- タグで記述される
 - ネストできる

```
<name>
```

```
  <surname>時実</surname>
```

```
  <given-name>象一</given-name>
```

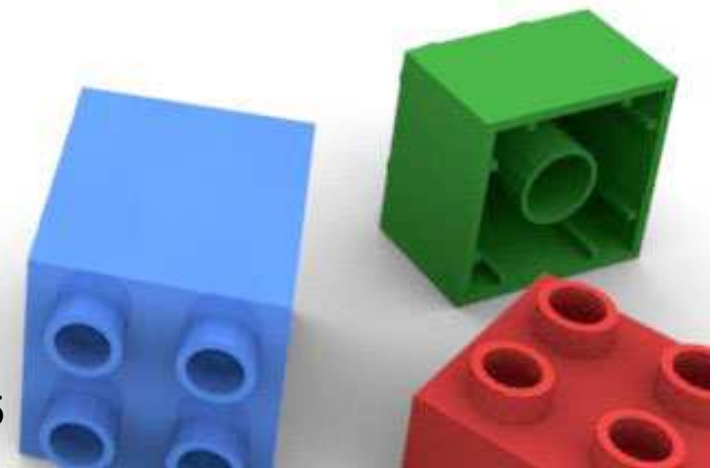
```
</name>
```



属性 (Attribute)

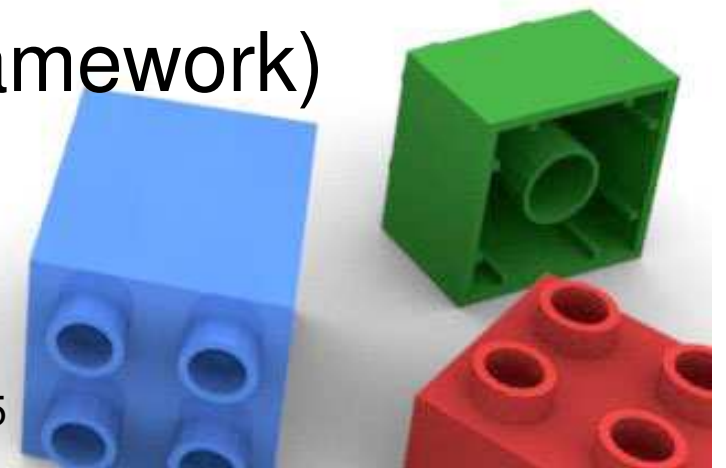
- 物の「性質」を表す
- 要素の付加的情報

`<name xml:lang="ja">中西秀彦</name>`

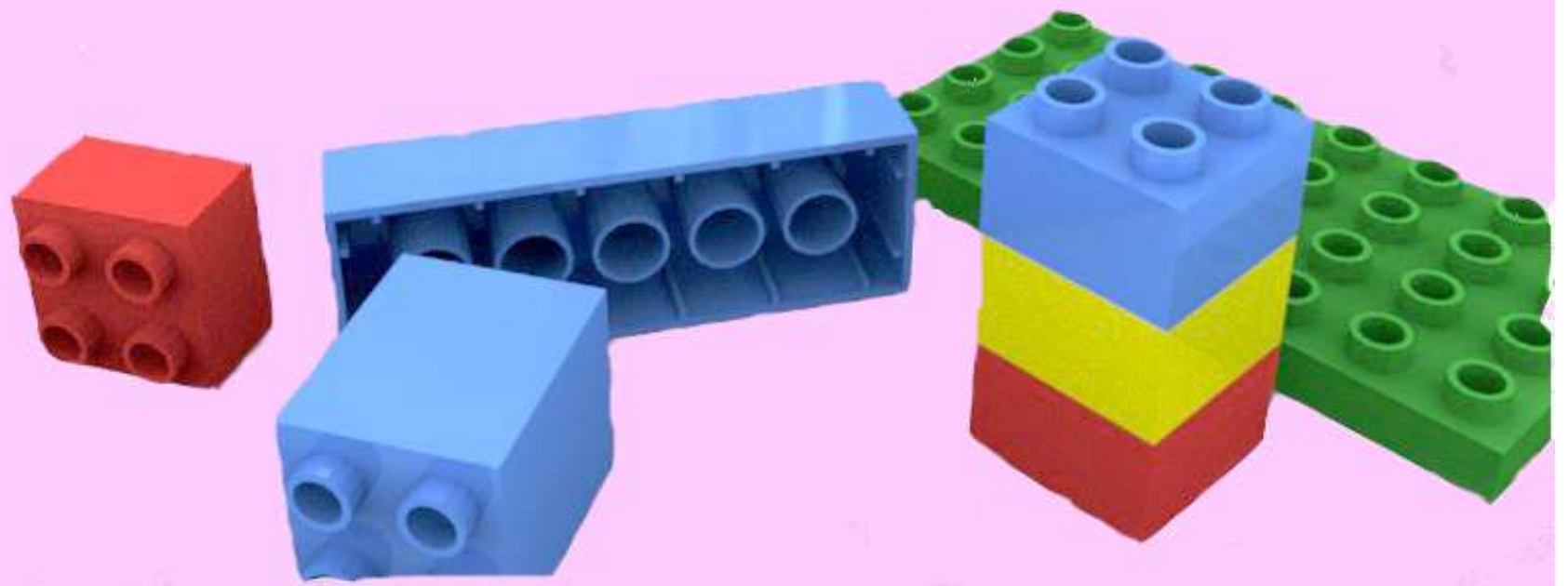


XMLの種類

- JATS ファミリー
 - JATS 学術雑誌記事記述
 - BITS 書籍記述
 - STS 規格文書 (ISO など) 記述
- JapaX 書籍用
- TEI (Text Encoding Initiative)
 - (学術) 文書の構造記述
- RDF (Resource Description Framework)
 - メタデータの記述
 - Linked Open Data (LOD) で

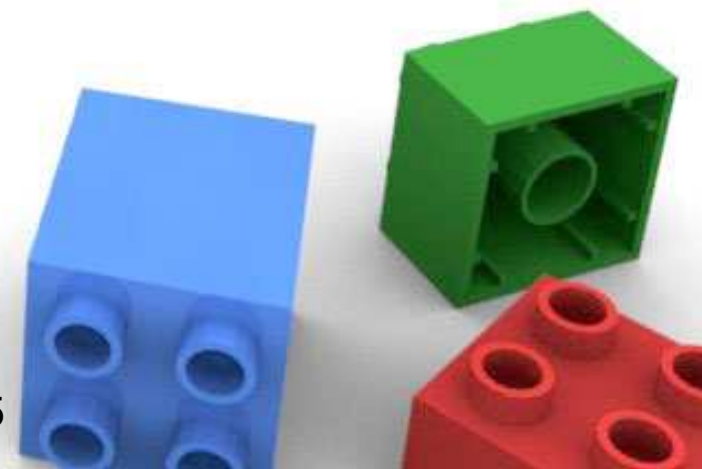


JATS



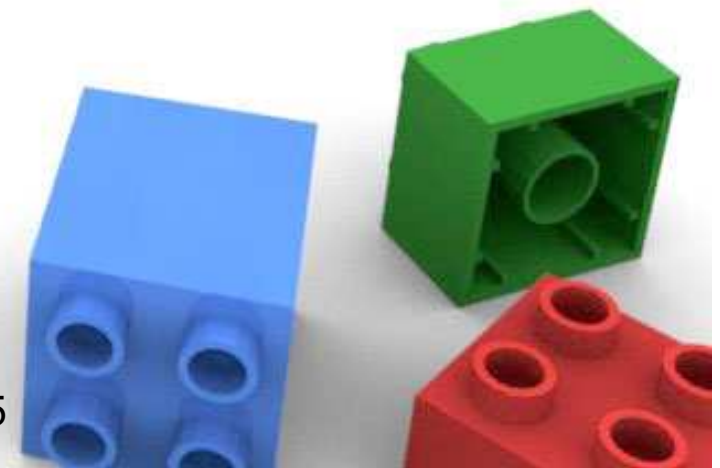
JATS

- Journal Article Tag Suite (JATS)
 - 米国国立医学図書館が開発した NLM DTD が基礎
 - NISO (National Information Standards Organization) の規格
 - 学術論文のアーカイブと流通・出版のための規格
 - 現在世界の学術論文出版の共通規格となっている
 - 日本からの提案を基礎として多言語に対応



JATS Family

タグセット	説明
Archiving and Interchange Tag Set	学術論文のアーカイブのためのタグセット
Journal Publishing Tag Set	学術論文データの交換と出版のためのタグセット
Article Authoring Tag Set	新規な雑誌記事執筆のためのタグセット
BITS (Book Interchange Tag Suite)	JATS を書籍のタグ付けに拡張
STS (Standards Tag Suite)	規格文書用のタグセット



JATS 1.1 翻訳版

<http://ejournal.jp/Links/StandardsCodes/JATS-Japanese/index.html>

- How To Use (Read Me First)
- Root Element
- General Introduction
- Selecting a Model & Schema
- ▶ Elements
- ▶ Attributes
- ▶ Parameter Entities
- ▶ Document Hierarchy Diagrams
- Full Article Samples
- ▶ Common Tagging Practice
- Accessibility
- ▶ Modifying This Tag Set
- Version 1.1 Change Report
- ▶ Element Context Table
- ▶ Index
- Supporting Documentation Home

Journal Publishing Tag Library NISO JATS Version 1.1 (ANSI/NISO Z39.96-2015)

December 2015

日本語翻訳版 (June 2018)

学術情報XML推進協議会 - XSPA JATS 規格検討分科会

*National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)*

Developer:

Mulberry Technologies, Inc.
17 West Jefferson Street, Suite 207
Rockville, MD 20850
Phone: 301/315-9631
Fax: 301/315-8285

<aff> Affiliation (所属先)

寄与者が所属する機関または団体名 (例えば大学、会社など)。

備考

1つの <aff> 要素内に複数層組織のすべてのレベルがリストされます。例えば、プログラム、学科、大学は同じ <aff> の一部であるかもしれません。同様に、部門と法人の両方が1つの <aff> 内にあることがあります。

推奨: 典型的な場合、ある <aff> 要素の @id 属性は <aff> 1つ以上の <contrib> 要素から指し示され、それらの間のリンクが確立されます。

明示的にタグ付けされた著者リンケージの番号または記号が変換時に保存されていても、変換中に ID/IDREF ID / IDREFリンク機構を維持または作成する必要があります。

関連要素

所属先要素 (<aff>) 内で機関名はテキストとして与えられるだけでもよいし、機関に明示的に (<institution>) としてタグ付けされ、機関識別子 (<institution-id>) が記録されてもよい。要素 <institution-wrap> は、機関名と識別子の両方を収集します。

異なる識別子機関が独自に識別子を定義しているか、またはそのエンティティが複数のより大きい組織単位に属しているため、単一の機関が複数の識別子を持つことがあります。たとえば、研究所が複数の機関の一部である可能性があります。

属性

content-type (コンテンツ種別)
 id (文書の内部識別子)
 rid (識別子への参照)
 specific-use (特定用途)
 xml:base (基底)
 xml:lang (言語)

コンテンツ・モデル

```
<!ELEMENT aff (#PCDATA %aff-elements;)* >
```

How To Use (Read Me First)

Root Element

General Introduction

Selecting a Model & Schema

▼ Elements

<abbrev>

<abbrev-journal-title>

<abstract>

<access-date>

<ack>

<addr-line>

<address>

<aff>

<aff-alternatives>

<ali:free_to_read>

<ali:license_ref>

<alt-text>

<alt-title>

<alternatives>

<annotation>

<anonymous>

<app>

<app-group>

<array>

<article>

<article-categories>

<article-id>

<article-meta>

<article-title>

<attrib>

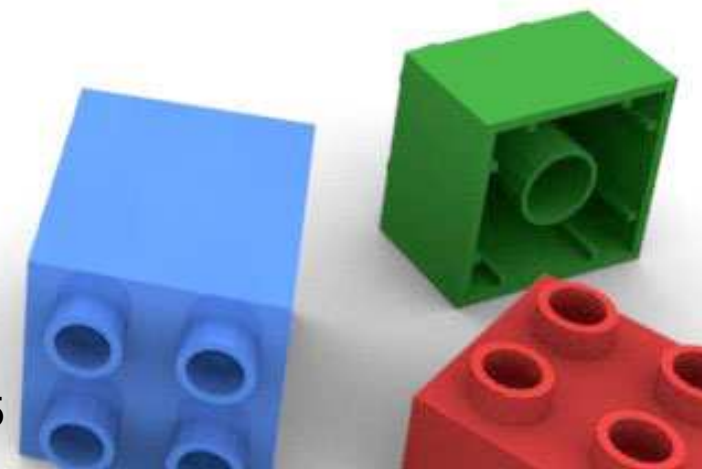
<author-comment>

<author-notes>

<award-group>

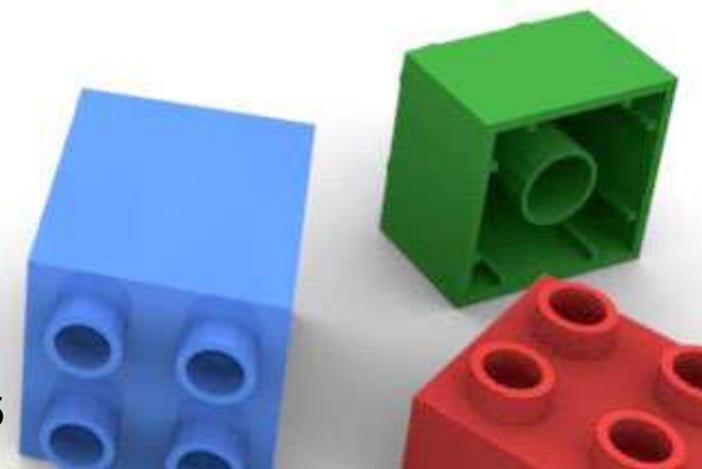
JATS の特徴

- 学術論文記事の記述に特化
 - 当初は科学技術医学が中心
 - 人文・社会科学にも適用が広がる
 - 日本語を含め多言語に対応
- 記事単位の XML ファイル
 - 雑誌単位の情報 (目次や巻末索引など) はない



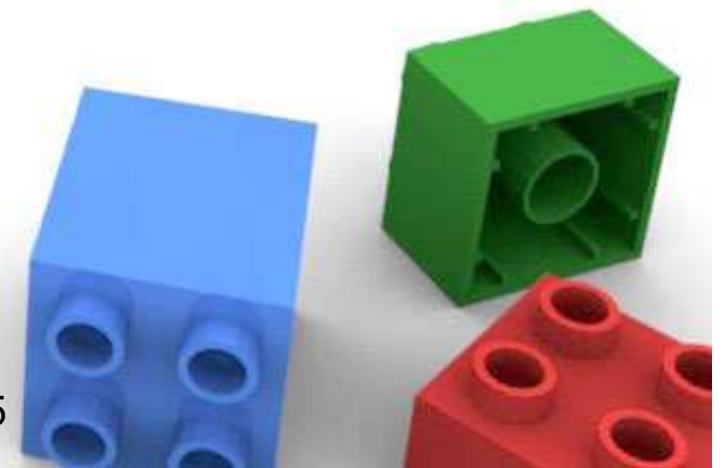
JATS (NLM DTD) の利用状況 (欧米)

1. 外部との流通にのみ利用
 - 大手出版社・大手学会など
 - 社内では独自の XML を利用
 - 外注のインターフェースは JATS が増加
2. JATS を拡張して社内でも利用
 - 自社独自の要素を追加
3. JATS の一部を社内でも利用



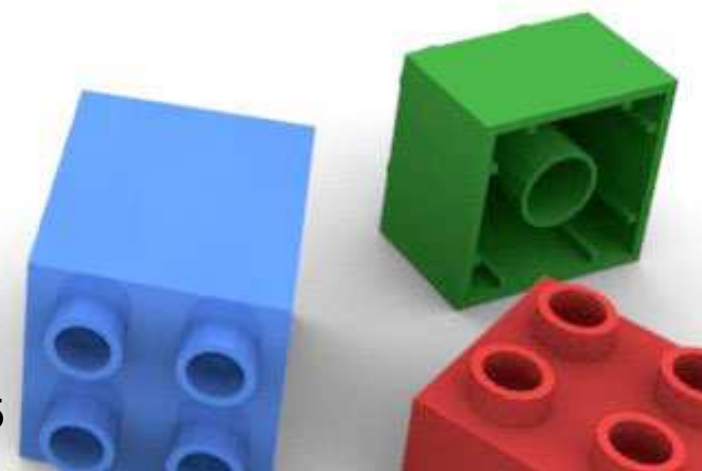
JATS の利用状況 (日本)

- 科学技術振興機構 (JST) の電子ジャーナルサイト J-STAGE で採用
- 医学中央雑誌で採用
- 「医書.jp オールアクセス」 (電子ジャーナルサービス) で採用



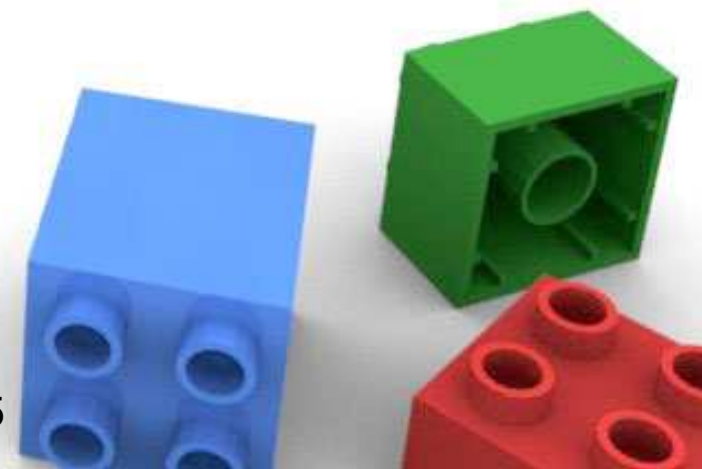
JATS の構成

- 構成要素
 - 宣言部分
 - 記事 (Article)
 - 前付け (Front matter)
 - 本文 (Body)
 - 後付け (Back matter)
 - 浮動部分 (Floating materials)



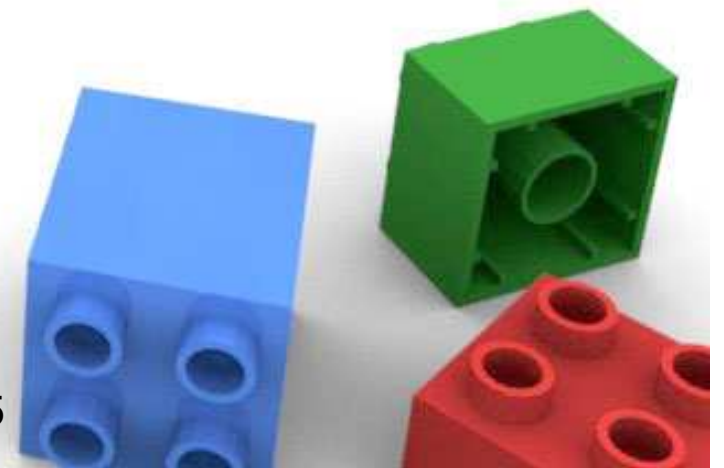
前付け

- <front>
 - <journal-meta> (雑誌メタデータ)
 - <article-meta> (記事メタデータ)
 - <article-id> (識別子)
 - <article-title> (記事タイトル)
 - <volume> (巻)
 - <issue> (号)
 - <fpage> (先頭ページ)
 - <lpage> (最終ページ)
 - <abstract> (抄録)
 - <contrib> (寄与者)
 - <name> (個人名)
 - <aff> (所属先)



<name>

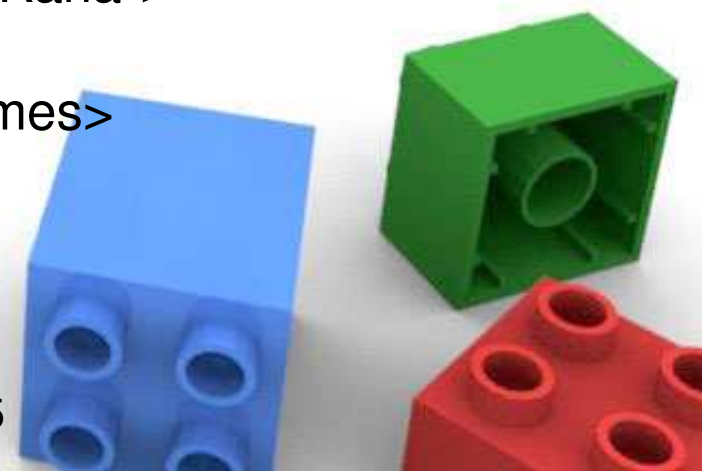
- 人名を記述する
 - <surname>, <given-name> に分けることができる
- ```
<name>
 <surname>Llanos De La Torre Quiralte</surname>
 <given-names>M</given-names>
</name>
```



# <name-alternatives>

- 同一人物に複数の名前がある場合に用いる
- 複数の <name> を含むことができる

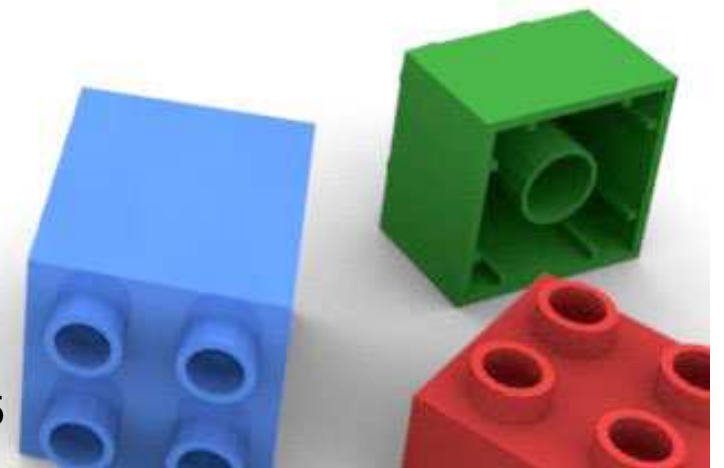
```
<name-alternatives>
 <name name-style="eastern" xml:lang="ja-Jpan">
 <surname>中西</surname>
 <given-names>秀彦</given-names>
 </name>
 <name name-style="western" xml:lang="en">
 <surname>Nakanishi</surname>
 <given-names>Hidehiko</given-names>
 </name>
 <name name-style="eastern" xml:lang="ja-Kana">
 <surname>ナカニシ</surname>
 <given-names>ヒデヒコ</given-names>
 </name>
</name-alternatives>
```





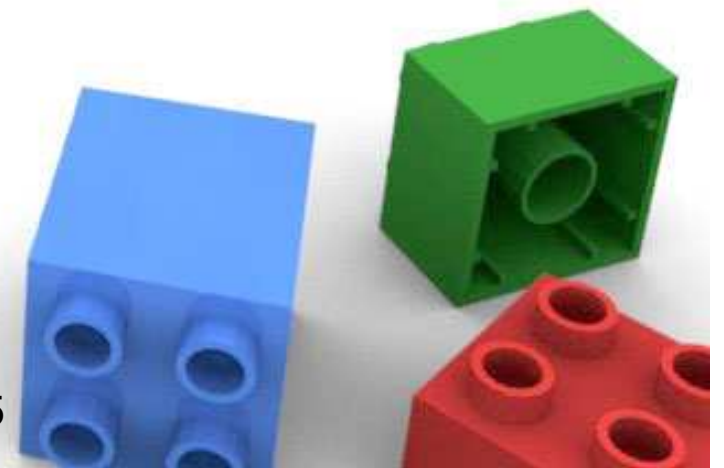
# 本文

- `<body>`
  - `<sec>` (セクション)
    - `<title>` (タイトル)
    - `<p>` (パラグラフ)
    - `<fig>` (図)
    - `<table>` (表)
    - `<fn>` (脚注)



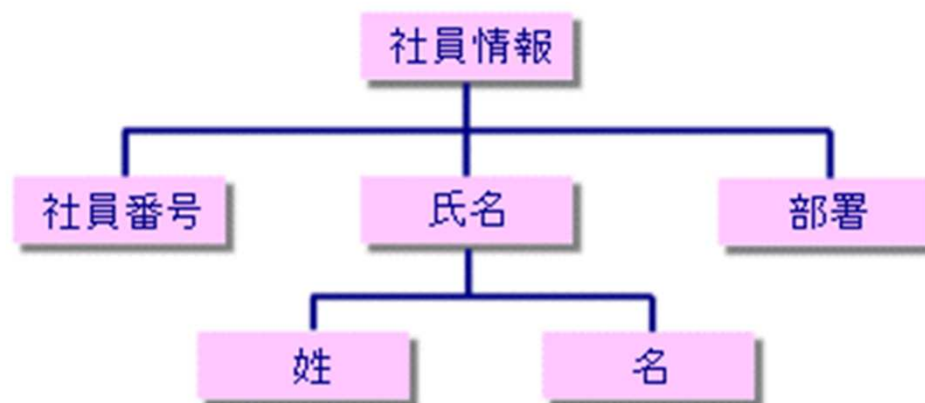
# 後付け

- <back>
  - <ack> (謝辞)
  - <ref-list> (文献リスト)
  - <notes> (注記)

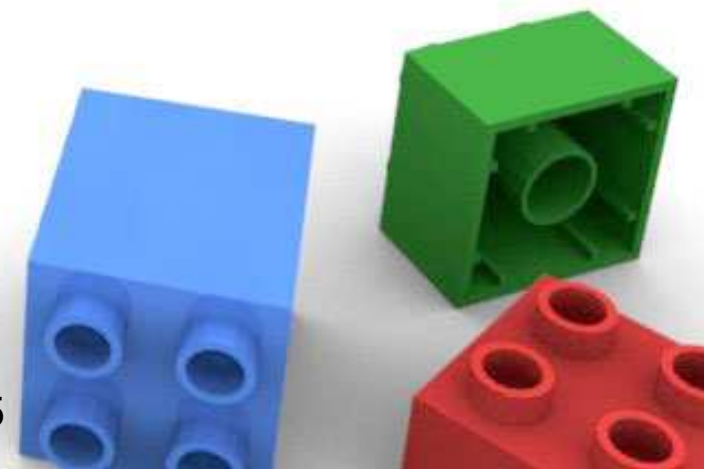


# なぜ JATS XML が必要か

- 雑誌記事の「構造」と「スタイル」を分離
- 構造

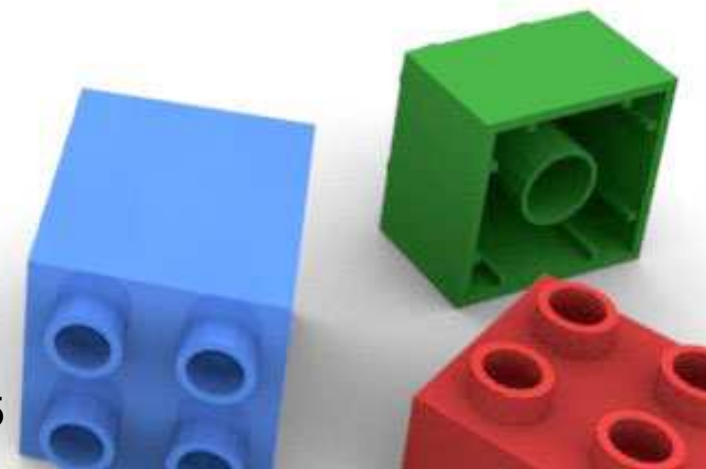


- スタイル
  - フォントやページ上の配置

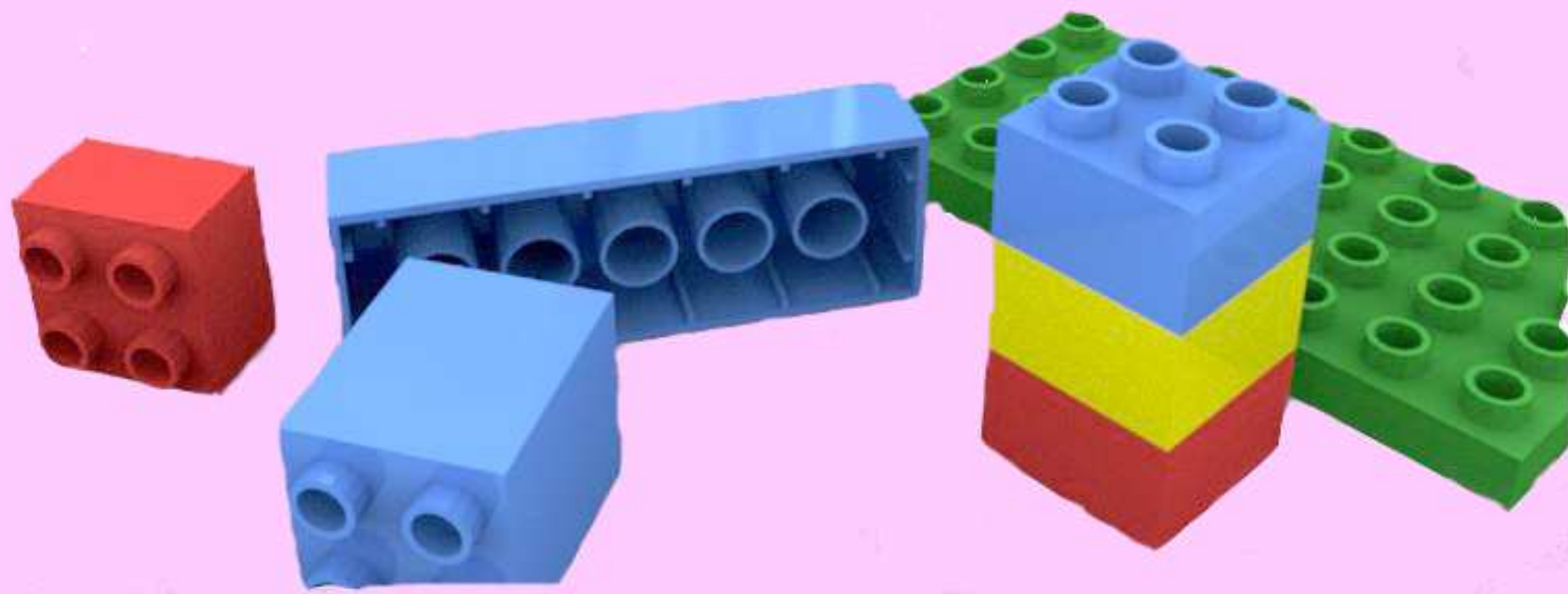


# なぜ JATS XML が必要か

- 構造だけを JATS として保存すると
  - One-Source Multi-Use が可能となる
  - ひとつの記事データを多種類出版できる
    - 印刷、PDF、HTML、EPUB
  - マイクロコンテンツの利用が可能となる
    - 特定のパラグラフ、図表など

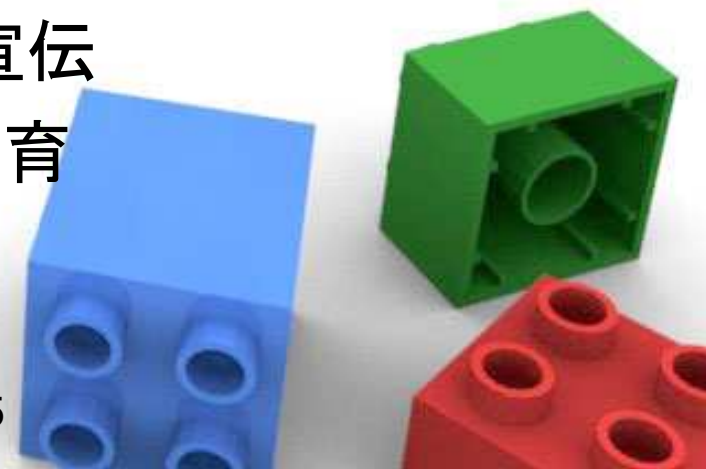


# JATS と XSPA (学術情報 XML 協議会)

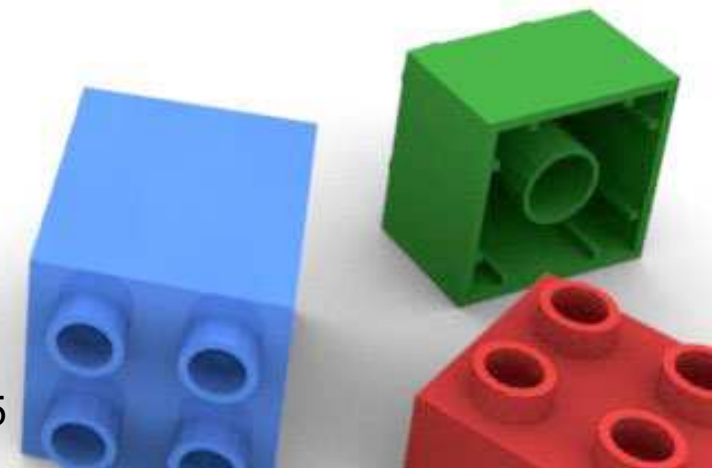


# 学術情報 XML 推進協議会 (XSPA)

- 学術情報の XML 製作・流通を推進するため  
結成 (2012/6/28)
- 前身は SPJ (Scholarly Publishing Japan)
  - 学会、印刷会社、JST、学識者
  - NLM DTD の日本語対応の研究と提言
  - JATS として実現
- 活動
  - XML 普及のための活動、広報・宣伝
  - 学会・印刷会社への情報提供・教育
  - JATS の改良のための研究・助言



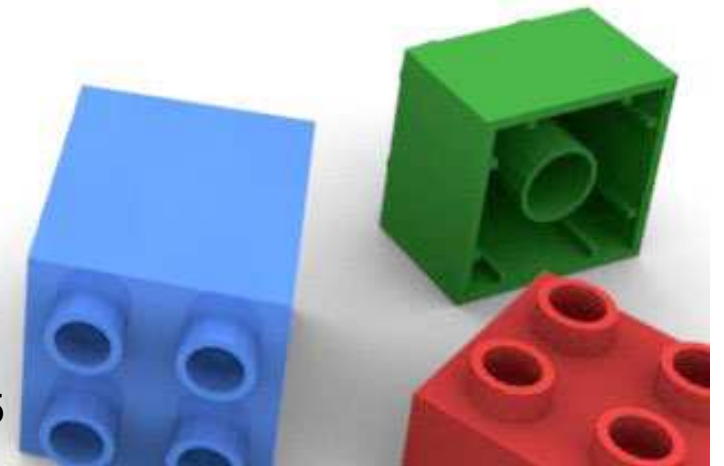
- 国際的な貢献
  - NISO JATS Steering Committee に時実が参加
  - JATS 誕生の際、日本語処理について提案
    - 現在の多言語化の基礎
  - ルビの導入、圏点の導入について提案
    - 実現
  - 縦書きについて提案予定



# NML DTD 時代の多言語記述例 (米国物理学会)

```
<author affref="a1 a2">
 <givenname>Haozhao</givenname>
 <surname>Liang</surname>
 <native-name lang="chitdr">梁豪兆</native-name>
</author>
```

- 3以上のバリエーションがあった場合対応が面倒

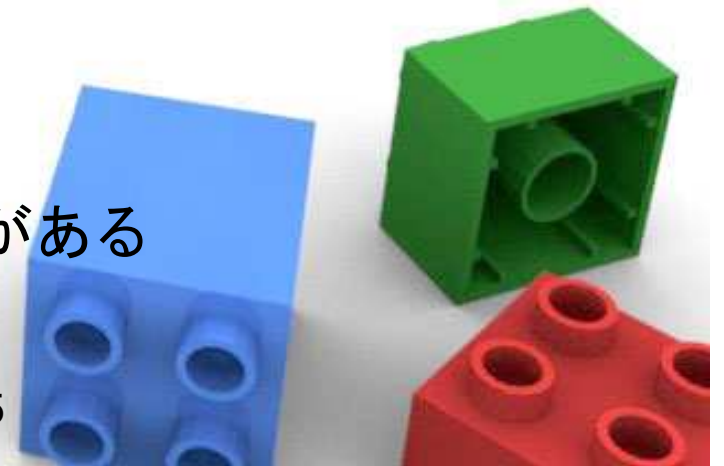




# NML DTD 時代の多言語記述例 (Atypon)

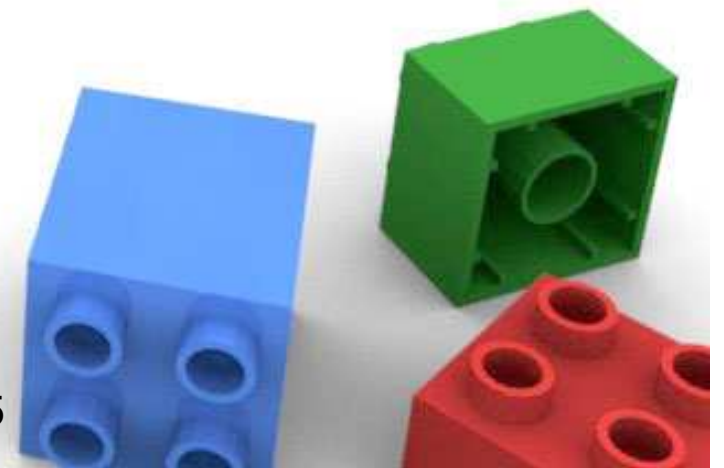
```
<contrib-group>
 <contrib contrib-type="author">
 <name name-style="western">
 <surname>Nihon</surname>
 <given-names>Taro</given-names>
 </name>
 <name name-style="eastern">
 <surname>日本</surname>
 <given-names>太郎</given-names>
 </name>
 </contrib>
</contrib-group>
```

- "eastern", "western" で言語を示すのは無理がある

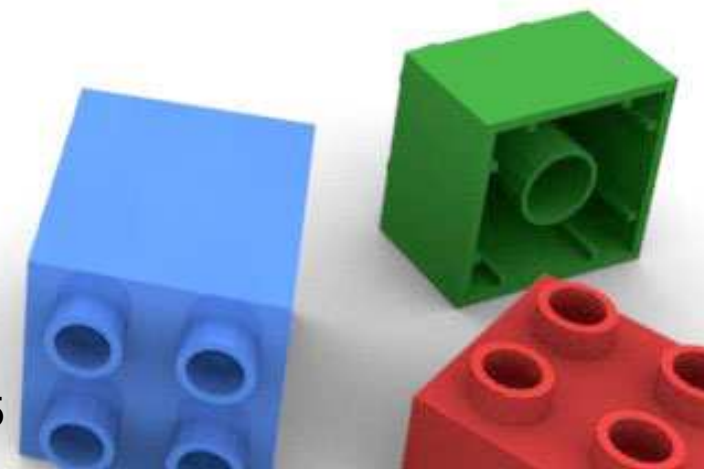


# JATS の多言語記述

- @name-style は姓名の配置順序を表す
- @xml:lang は名前の言語とスクリプトを表す



```
<contrib-group>
<contrib contrib-type="author">
 <name name-style="eastern" xml:lang="ja-Jpan">
 <surname>中西</surname>
 <given-names>秀彦</given-names>
 </name>
 <name name-style="western" xml:lang="en">
 <surname>Nakanishi</surname>
 <given-names>Hidehiko</given-names>
 </name>
 <name name-style="eastern" xml:lang="ja-Kana">
 <surname>ナカニシ</surname>
 <given-names>ヒデヒコ</given-names>
 </name>
</contrib>
</contrib-group>
```



# JATS のルビ

## 用例 1

畿内や西日本一帯では<sup>むぎ</sup>麦を<sup>うらさく</sup>裏作とする二毛作が普及していった。

```
...
<p>... 畿内や西日本一帯では<ruby><rb>麦</rb>
<rt>むぎ</rt></ruby>を<ruby><rb>裏</rb><rt>うら</rt></ruby>
<ruby><rb>作</rb><rt>さく</rt></ruby>とする二毛作が普及していった。
...</p>
...
```

## 用例 2

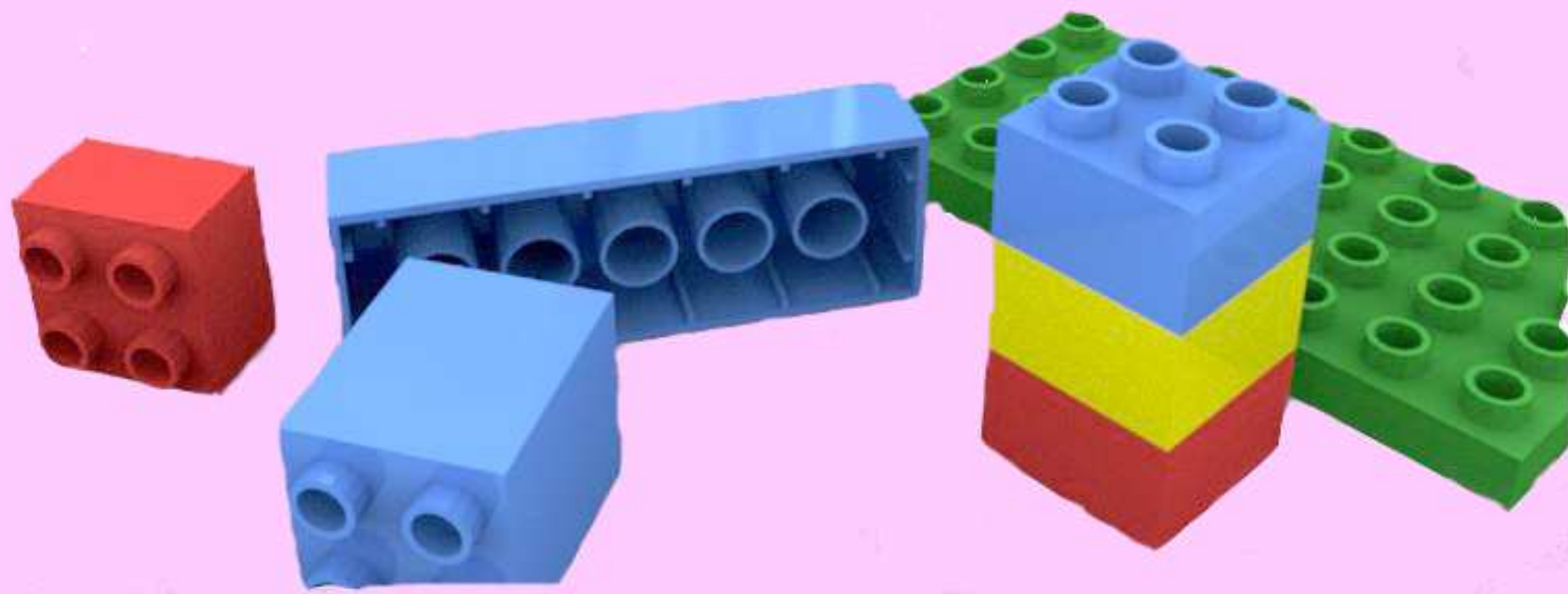
<sup>Scotland Yard</sup>  
あのロンドン警視庁は霧の中に隠れていた。

or

<sup>ろんどんけいしちょう</sup>  
あのロンドン警視庁は霧の中に隠れていた。

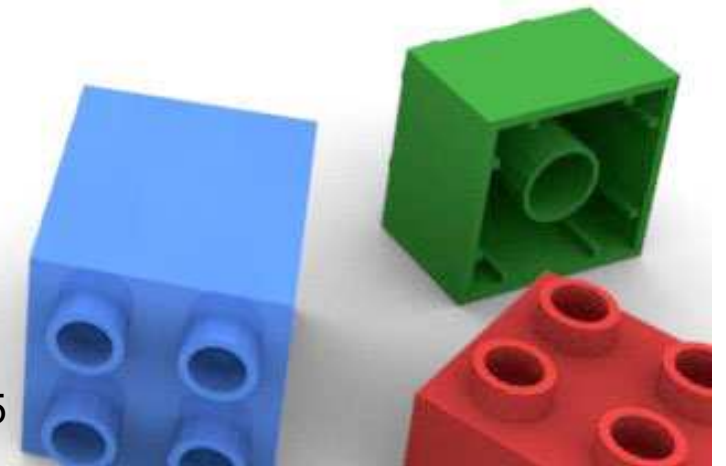
```
...
<p>... あの<ruby><rb>ロンドン警視庁</rb>
<rt xml:lang="en">Scotland Yard</rt></ruby>は霧の中に隠れていた。
or
あの<ruby><rb>ロンドン警視庁</rb>
<rt xml:lang="ja">ろんどんけいしちょう</rt>
</ruby>は霧の中に隠れていた。 ...</p>
...
```

# JATS の周辺



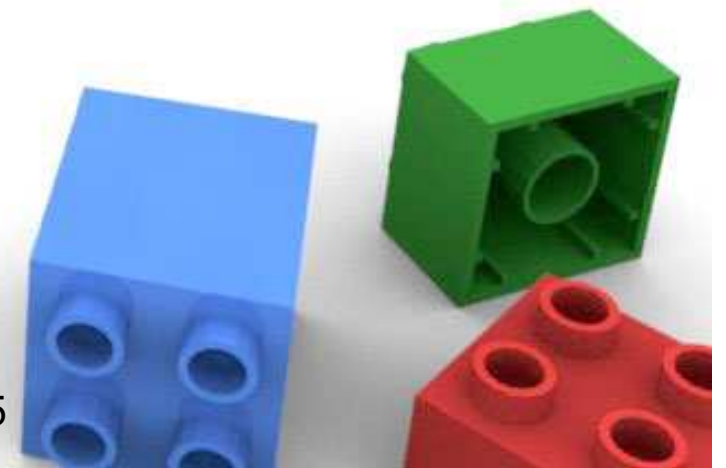
# JATS の周辺

- 規格文書
- BITS
- スキマトロンによる論理チェック
- 記事の自動 (半自動) 組版
- InDesign との連携
- 執筆ツール



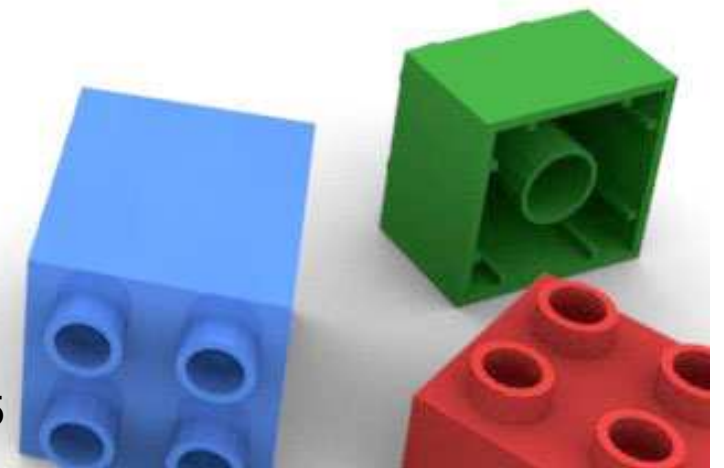
# 規格文書

- Standards Tag Suite (STS)
- ISO で採用する他、多くの各国規格機関で採用
  - JIS は未採用



# BITS

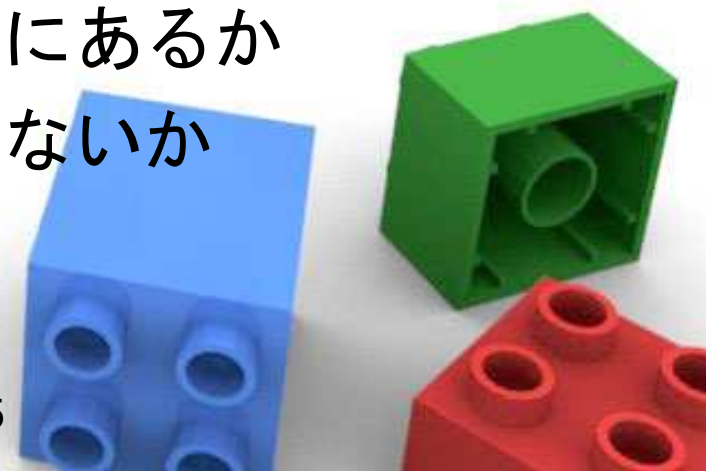
- 書籍のためのタグセット
- Book Interchange Tag Suite
  - <book-meta>
    - 書誌事項
  - <front-matter>
    - 前書き、序文、目次
  - <book-body>
    - 本文
    - <book-part>
      - 部、章、節など
  - <book-back>
    - あとがき、解説、索引





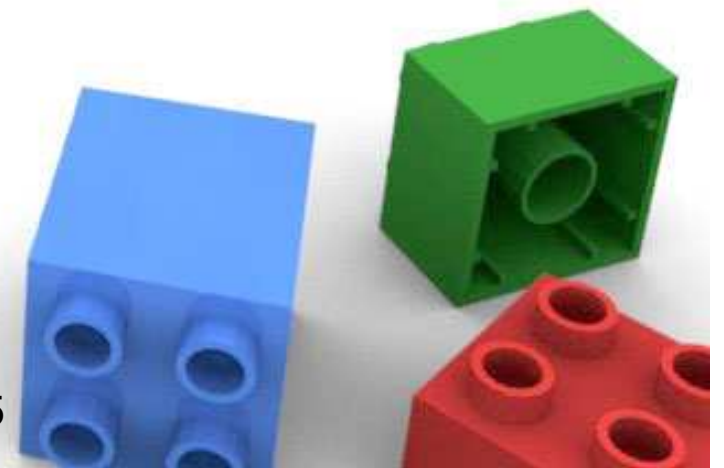
# スキマトロン

- XML データの論理チェックができる
- 例
  - 年月日が未来 (e.g. 2031年) になっていないか
  - 架空の日付はないか (e.g. 13月45日)
  - 最終ページ数が先頭ページ数より大きいか
  - 著者名が正しく姓と名からなっているか
  - 文献リストに対応した文献番号が本文中にあるか
  - 図表に対応した図表番号が本文中にあるか
  - 全角文字と半角文字が混ざっていないか



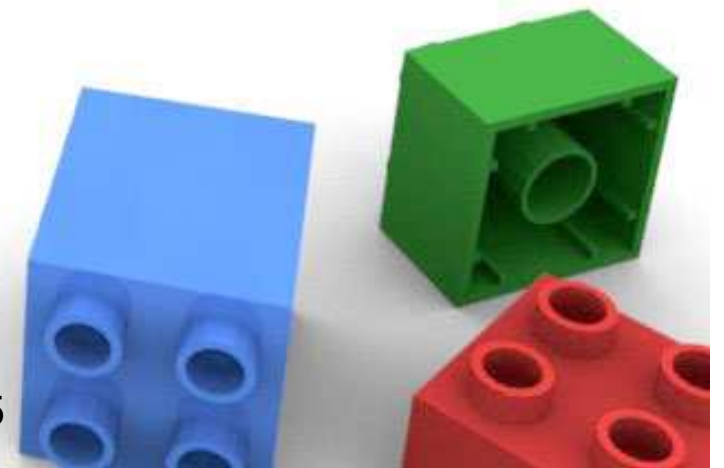
# 記事の自動 (半自動) 組版

- アンテナハウス AH Formatter
- その他
- PDF, EPUB を作成



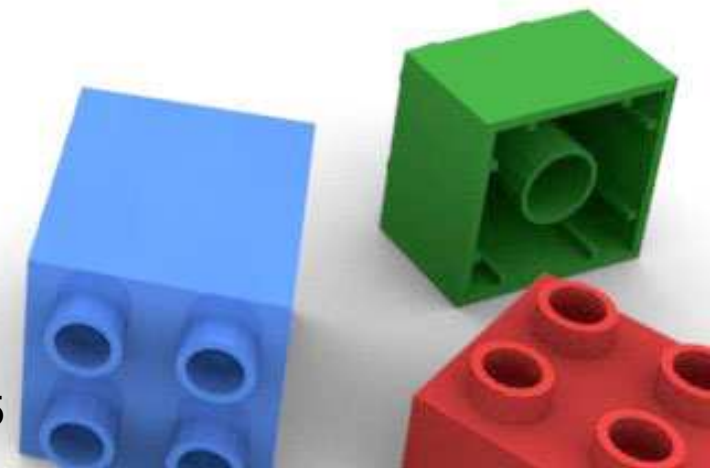
# InDesign との連携

- JATS から InDesign
- InDesign から JATS



# PDF または DOCX から作成

- J-STAGE 「全文 XML 作成ツール」



# 執筆ツール

- Authorea
- Manuscripts.io



## 紙の本はなくなるのか

時実 象一

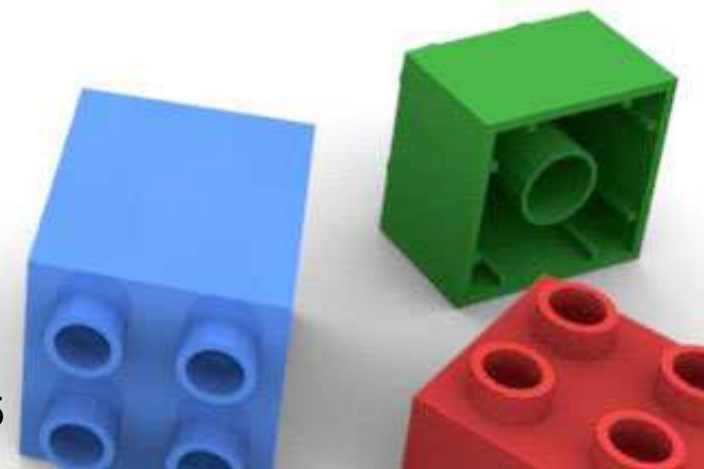
### 1. 電子書籍とは

電子書籍とは、電子的に読める本のことで、はじめから電子的に作成されたもの (born digital) と冊子体が電子化されたものがある。



# XML, JATS の利点

- 構造化された記事文書が作成できる
  - 文書の要素がタグ付けされている
- 記事文書の多角的利用が可能
  - Create Once, Publish Everywhere
  - PDF, EPUB, Web
- 記事の国際的流通



- 時実 象一
- ときざね そういち
- [tokizane@pc.highway.ne.jp](mailto:tokizane@pc.highway.ne.jp)
- <http://tokizane.jp/>
- ORCID: <http://orcid.org/0000-0003-1236-1930>

