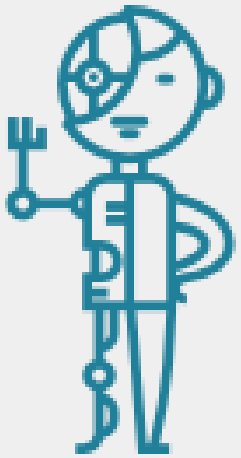
INERA
eXtyles | edifix

# Finding the Tipping Point in Automated Markup During Up-Translation

*[or, Cyborgs are the Future of Scholarly Publishing]*

**Caitlin Gebhard**
Inera Inc.



# What are we talking about?

- Up-translation for scholarly publishing
- Limits of automated XML conversion
- Semi-automated conversion techniques

# Up-Translation for Scholarly Publishing

Authors,* editors

Editorial/
production magic

Production,
publishing ecosystem

* LaTeX *is really cool but out of scope today.*

INERA
eXtyles | edifix

# XML drives:

- Submission and peer review systems
- Digital object identifiers (DOIs)
- ORCID for contributor identifiers
- FundRef data for funding information
- Accessibility initiatives
- Single-source publishing to Web, E-pub, PDF, print, and beyond!

# Journal Article Tag Suite (JATS)

- **2003**: NLM DTDs published by NCBI (NLM)
- **2008**: NLM DTD submitted to NISO
- **2012**: “JATS” standardized as ANSI/NISO Z39.96
- **2015**: JATS 1.1 released

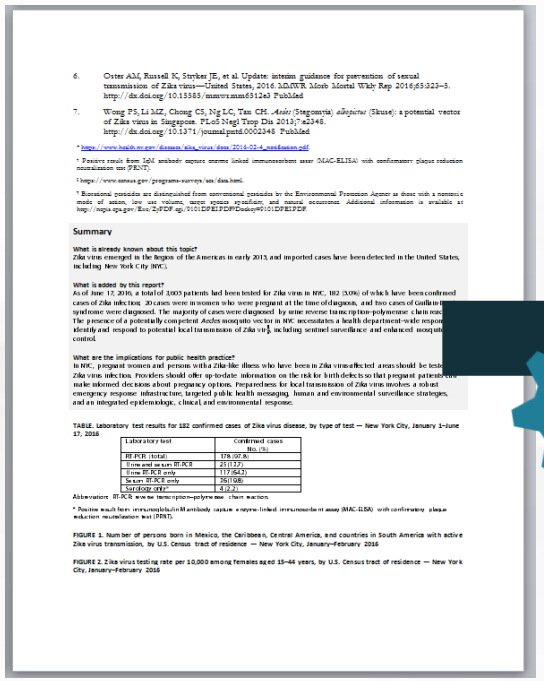
# Why JATS XML?

- Pros:
  - Industry standard
  - Highly-structured, rich, and granular
  - Complex and flexible

# “Cons?”

- Because JATS is rich and complex:
  - 100% *automated* conversion is not realistic
  - 100% *manual* conversion is not realistic

# Word Doc → JATS XML



6. Oster AM, Russell K, Stryker JE, et al. Update: interim guidance for prevention of sexual transmission of Zika virus—United States, 2016. MMWR Morb Mortal Wkly Rep 2016;65:323–5. http://dx.doi.org/10.15585/mmwr.mm6512e3 PubMed

7. Wong PS, Li MZ, Chong CS, Ng LC, Tan CH. *Aedes* (Stegomyia) *albopictus* (Skuse): a potential vector of Zika virus in Singapore. PLoS Negl Trop Dis 2013;7:e2348. http://dx.doi.org/10.1371/journal.pntd.0002348 PubMed

* https://www.health.ny.gov/diseases/zika_virus/docs/2016-02-4_nov6assay.pdf

† Possible result from IgM antibody capture enzyme linked immunosorbent assay (MAC-ELISA) with confirmatory plaque reduction neutralization test (PRNT).

§ https://www.census.gov/programs-surveys/acs/data.html.

¶ Biorational pesticides are distinguished from conventional pesticides by the Environmental Protection Agency as those with a nontoxic mode of action, low use volume, target species specificity, and natural occurrence. Additional information is available at http://nepis.epa.gov/Exe/ZyPDF.cgi/P1010DPE.PDF?Dockey=P1010DPE.PDF.

**Summary**

**What is already known about this topic?**
Zika virus emerged in the Region of the Americas in early 2015, and imported cases have been detected in the United States, including New York City (NYC).

**What is added by this report?**
As of June 17, 2016, a total of 1,003 patients had been tested for Zika virus in NYC, 182 (3.0%) of which have been confirmed cases of Zika infection; 20 cases were in women who were pregnant at the time of diagnosis, and two cases of Guillain-Barré syndrome were diagnosed. The majority of cases were diagnosed by urine reverse transcription-polymerase chain reaction. The presence of a potentially competent *Aedes* mosquito vector in NYC necessitates a health department-wide response to identify and respond to potential local transmission of Zika virus, including sentinel surveillance and enhanced mosquito control.

**What are the implications for public health practice?**
In NYC, pregnant women and persons with a Zika-like illness who have been in Zika virus–affected areas should be tested for Zika virus infection. Providers should offer up-to-date information on the risk for birth defects so that pregnant patients can make informed decisions about pregnancy options. Preparedness for local transmission of Zika virus involves a robust emergency response infrastructure, targeted public health messaging, human and environmental surveillance strategies, and an integrated epidemiologic, clinical, and environmental response.

TABLE. Laboratory test results for 182 confirmed cases of Zika virus disease, by type of test — New York City, January 1–June 17, 2016

| Laboratory test | Confirmed cases No. (%) |
|---|---|
| RT-PCR (total) | 176 (97.8) |
| Urine and serum RT-PCR | 57 (32) |
| Urine RT-PCR only | 117 (66) |
| Serum RT-PCR only | 2 (1) |
| Serology only* | 4 (2.2) |

Abbreviation: RT-PCR = reverse transcription–polymerase chain reaction.

* Positive result from immunoglobulin M antibody capture enzyme-linked immunosorbent assay (MAC-ELISA) with confirmatory plaque reduction neutralization test (PRNT).

FIGURE 1. Number of persons born in Mexico, the Caribbean, Central America, and countries in South America with active Zika virus transmission, by U.S. Census tract of residence — New York City, January–February 2016

FIGURE 2. Zika virus testing rate per 10,000 among females aged 15–44 years, by U.S. Census tract of residence — New York City, January–February 2016

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE article PUBLIC "-//NLM//DTD JATS (Z39.96) Journal Archiving and Interchange DTD v1.0 20120330//
<article article-type="research-article" dtd-version="1.0" xml:lang="en" xmlns:xlink="http://www.w3.org/1
<front>
<journal-meta>
<journal-id journal-id-type="publisher-id">SS</journal-id>
<journal-id journal-id-type="nlm-ta">MMWR Surveill Summ</journal-id>
<journal-title-group>
<journal-title>MMWR. Surveillance Summaries</journal-title>
<abbrev-journal-title abbrev-type="pubmed">MMWR Surveill. Summ.</abbrev-journal-title>
</journal-title-group>
<iss pub-type="ppub">1546-0738</issn><issn pub-type="epub">1545-8636</issn>
<her><publisher-name>Centers for Disease Control and Prevention</publisher-name></publisher>
eta>
ta>
<arti pub-id-type ublisher-id">123</article-id>
<artic b-id-type 0.15585/</article-id>
-group-type="heading"><subject>Surveillance Summaries</subject></subj
ble Diseases &#x2014; United States, 2012</article-title>
group>
ontrib-group>
<contrib contrib-type="author"><name><surname>Adams</surname><given-names>Deborah A.</given-names></name>
</contrib-group>
<contrib-group>
<contrib contrib-type="author"><name><surname>Ajani</surname><given-names>Umed</given-names></name><degre
<contrib contrib-type="author"><name><surname>Jajosky</surname><given-names>Ruth Ann</given-names></name>
<contrib contrib-type="author"><name><surname>Kriseman</surname><given-names>Jeffrey</given-names></name>
<contrib contrib-type="author"><name><surname>Sharp</surname><given-names>Pearl</given-names></name></con
<contrib contrib-type="author"><name><surname>Onweh</surname><given-names>Diana H.</given-names></name></
<contrib contrib-type="author"><name><surname>Schley</surname><given-names>Alan W.</given-names></name></
<contrib contrib-type="author"><name><surname>Anderson</surname><given-names>Willie J.</given-names></nam
<contrib contrib-type="author"><name><surname>Grigoryan</surname><given-names>Anna</given-names></name><d
<contrib contrib-type="author"><name><surname>Aranas</surname><given-names>Aaron E.</given-names></name><
<contrib contrib-type="author"><name><surname>Wodajo</surname><given-names>Michael S.</given-names></name
```

**Input** :
varied, wild,
made by humans

**Output** :
highly structured,
granular, complex

INERA
eXtyles | edifix

# Automated Tagging

*"**During the RFP process, we heard statements such as 'we can provide 100% accuracy and 100% automation'. This is simply impossible,** and Innodata were one of the more open providers initially offering 98% accuracy and being quite upfront this would be a semi-automated process, ie, a large part of the operation would require manual tagging."*

Galichet, L. "Beware of the Laughing Horse: Managing a Back-catalogue Conversion;" JATS Con 2017.

INERA
eXtyles | edifix

# Manual Tagging

- "Pros:"
  - (Seems) inexpensive
- Cons:
  - *Time*
  - Inconsistencies
  - One-off errors

## Example

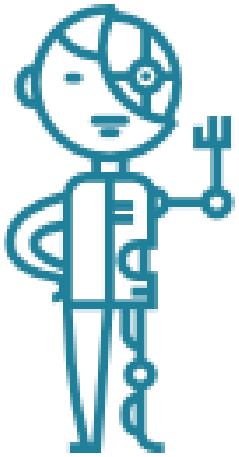
```
<mixed-citation
publication-type=
  "jornal".../>
```

# Manual Tagging

- "Pros:"
  - (Seems) inexpensive
- Cons:
  - *Time*
  - Inconsistencies
  - One-off errors

## Example

```
<mixed-citation
publication-type=
"journal".../>
```



Semi-
Automation for
**Metadata**

# Contributor Metadata

Bad tagging means:

- Overlooked credit or misattribution of credit
- Incorrect references to the author that appear in subsequent papers
- Confusion at funding organizations
- Unhappy authors

# Bibliographic References Metadata

Bad tagging means:

- Missing or incorrect reference links
- Readers can't find cited source
- Overlooked credit or misattribution of credit
- Negative impact on cited publication's measurability and discoverability
- Unhappy authors *and* researchers

# **Limits of Automation:** Contributors

# Pattern Recognition 101

D Carol Bird MD, editor

INERA
eXtyles | edifix

# Pattern Recognition 101

## D Carol Bird MD, editor

```
<contrib>
<name>
<surname>Bird</surname>
<given-names>D Carol</given-names>
</name>
<degrees>MD</degrees>
<role>editor</role>
</contrib>
```

INERA
eXtyles | edifix

# Machine vs Human

- Really fast
- Consistent

- Common sense
- Flexible
- Self-aware
- Can use Google
- Can ask the author

# Example 1:
# Name Segmentation

Jonathan Taylor Thomas

```
<contrib contrib-type="author">
<name>
<surname>Thomas</surname>
<given-names>Jonathan Taylor</given-names>
</name>
</contrib>
```

INERA
eXtyles | edifix

# Example 1:
# Name Segmentation

Federico García Lorca

```
<contrib contrib-type="author">
<name>
<surname>Lorca</surname>
<given-names>Federico Garc&#x00ED;a</given-names>
</name>
</contrib>
```

INERA
eXtyles | edifix

# Example 1:
# Name Segmentation

Federico García Lorca

```
<contrib contrib-type="author">
<name>
<surname>Garc&#x00ED;a Lorca</surname>
<given-names>Federico</given-names>
</name>
</contrib>
```

INERA
eXtyles | edifix

# Example 1:
# Name Segmentation

María Cristina Esquivel de Sánchez

```
<contrib contrib-type="author">
<name>
<surname>de S&#x00E1;nchez</surname>
<given-names>Mar&#x00ED;a Cristina Esquivel
</given-names>
</name>
</contrib>
```

# Example 1:
# Name Segmentation

## María Cristina Esquivel de Sánchez

```
<contrib contrib-type="author">
<name>
<surname>Esquivel de S&#x00E1;nchez</surname>
<given-names>Mar&#x00ED;a Cristina</given-names>
</name>
</contrib>
```

INERA
eXtyles | edifix

# Example 1:
# Name Segmentation

✗ Charles Maurice de Talleyrand-Périgord

✓ Charles Maurice de Talleyrand-Périgord

INERA
eXtyles | edifix

# Example 2:
# Surname or Suffix

Edward Araujo Júnior

≠ Edward Araujo Jr.

```
<contrib contrib-type="author"><name>
<surname>Araujo</surname>
<given-names>Edward</given-names>
<suffix>J&#x00FA;nior</suffix>
</name></contrib>
```

INERA
eXtyles | edifix

# Example 2:
# Surname or Suffix

Edward Araujo Júnior

~~≠ Edward Araujo Jr.~~

```
<contrib contrib-type="author"><name>
<surname>Araujo J&#x00FA;nior</surname>
<given-names>Edward</given-names>
</name></contrib>
```

INERA
eXtyles | edifix

# Example 3:
# Eastern-style Name Order

Zhou De-an

```
<contrib contrib-type="author"><name>
<surname>De-an</surname>
<given-names>Zhou</given-names>
</name></contrib>
```

# Example 3:
# Eastern-style Name Order

Zhou De-an

```
<contrib contrib-type="author"><name>
<surname>Zhou</surname>
<given-names>De-an</given-names>
</name></contrib>
```

# Example 3: Eastern-style Name Order

- What if the author list includes both Eastern- and Western-order names?
  - e.g., Han Solo and Han Sungee

# Example 3: Eastern-style Name Order

- Some authors "Westernize" their names
  - Lee Choon Shil → Choon Shil Lee
- Editors (and co-authors!) sometimes "re-Westernize" names
  - Lee Choon Shil → Choon Shil Lee → Lee Choon Shil (×)

INERA
eXtyles | edifix

# Example 3.1: Giving Up

Jae-Hong Chang, Yong-Sik Youn

```
<contributors>
<person_name sequence="first"
contributor_role="author">
<surname>Jae-Hong Chang</surname>
</person_name>
<person_name sequence="additional"
contributor_role="author">
<surname>Yong-Sik Youn</surname>
</person_name>...
```

INERA
eXtyles | edifix

# Limits of Automation Part 2:

# Parsing references isn't fun

- JATS is very granular
- Hundreds of different editorial styles
- Inconsistent formatting
- Errors introduced by reference tools
- Typos
- Weird stuff

# Example 4:
# Immunoglobulin A

Watanabe A; Kitamura M; Shimizu M; **Immunoglobulin A** (IgA) with properties of both cryoglobulin and pyroglobulin. *Clinica Chimica Acta*. 1974;52(2):231–237.

INERA
eXtyles | edifix

# Example 4:
# Immunoglobulin A

```
...
<string-name><surname>Immunoglobulin</surname>
<given-names>A</given-names></string-name>
</person-group>. <article-title>(IgA) with
properties of both cryoglobulin and
pyroglobulin.</article-title>...
```

INERA
eXtyles | edifix

# Example 4: Immunoglobulin A

```
...</person-group>.
<article-title>Immunoglobulin A (IgA) with
properties of both cryoglobulin and
pyroglobulin.</article-title> ...
```

# Example 5:
# False Duplicates

Morcos MW, Al-Jallad H, Hamdy R. Comprehensive review of adipose stem cells and their implication in distraction osteogenesis and bone regeneration. *BioMed Research International*. **2015** Sep 13;**2015**.

- Volume? Issue? Date? Page number?

INERA
eXtyles | edifix

# Example 5:
# False Duplicates

```
...<article-title>Comprehensive review of
adipose stem cells and their implication in
distraction osteogenesis and bone
regeneration.</article-title> <source>BioMed
research international<source>. 2015 Sep
13;<year>2015</year>.</mixed-citation></ref>
```

INERA
eXtyles | edifix

# Example 5:
# False Duplicates

```
...<article-title>Comprehensive review of
adipose stem cells and their implication in
distraction osteogenesis and bone
regeneration</article-title>. <source>BioMed
research international.</source>
<year>2015</year> <month>Sep</month>
<day>13</day>;<volume>2015</volume>.</mixed-
citation></ref>
```

INERA
eXtyles | edifix

# Example 6:
# Vital Health Stat 10

Benson V, Marano MA. Current estimates from the National Health Interview Survey, 1992. Vital Health Stat 10 1994 Jan;(189):1-269.

- **Hint**: *Vital and Health Statistics. Series 10. Data from the National Health Survey* is a journal title.
- Databases FTW

INERA
eXtyles | edifix

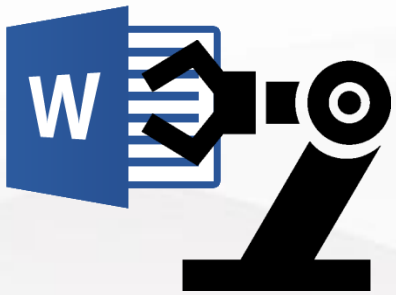
# Balancing Automation with Human Review

# Unassisted Automation

(Published online: 16 April 2009)

```
<pub-date pub-type="epub">
<day>16</day><month>04</month>
<year>2009</year></pub-date>
```

# Human-reviewed Automation



**Automatic parsing** → **Human review** → **Final conversion**

# Human-reviewed Automation

Automated application of **visual cues** for human review of machine-parsing

- Color-coding
- Unique character styles
- Quick and easy review

# Human-reviewed Automation

Jonathan Taylor Thomas and Federico García Lorca

Jonathan Taylor Thomas and Federico García Lorca

# Human-reviewed Automation

Kuck D, Caulfield T, Lyko F, Medina-Franco JL, **Nanaomycin A** Selectively Inhibits DNMT3B and Reactivates Silenced Tumor Suppressor Genes in Human Cancer Cells. Mol Cancer Ther 2010;9:3015–23.

Same issue as “Immunoglobulin A”

INERA
eXtyles | edifix

# Without proofing

Kuck D, Caulfield T, Lyko F, Medina-Franco JL, Nanaomycin A. Selectively Inhibits DNMT3B and Reactivates Silenced Tumor Suppressor Genes in Human Cancer Cells. Mol Cancer Ther 2010;9:3015–23.

# With proofing

<jrn>2.Kuck D, Caulfield T, Lyko F, Medina-Franco JL, Nanaomycin A. Selectively Inhibits DNMT3B and Reactivates Silenced Tumor Suppressor Genes in Human Cancer Cells. Mol Cancer Ther 2010;9:3015–23.</jrn>

INERA
eXtyles | edifix

# Flag suspect cases

- Automated insertion of **warnings** for human review of machine-parsing
- Help software recognize its limits
- Silent errors are not helpful

# Flag suspect cases

<jrn>57. Watanabe T, Kelsey L, Ageitos A, Kuszynski C, Ino Kk, Heimann DG et al. Enhancement of advenovirus-mediated gene transfer to human bone marrow cells. Leuk Lymphoma 1998; 29:439-51.</jrn>

**Comment [CG1]:** The process was unable to fully identify the author list in the reference, most likely because of incorrect punctuation in the original text. Please correct the author list by hand.

"The process was unable to fully identify the author list in the reference, most likely because of incorrect punctuation in the original text. Please correct the author list by hand."

INERA
eXtyles | edifix

# Post-Conversion Quality Assurance

- XSLT stylesheets to display a WYSIWYG version of the tagged content
  - Color and formatting
- Schematron for customized QA of XML

# Until AI editors are available…

- Converting human-created content into granular XML cannot be 100% automated with 100% accuracy
- 100% manual conversion is too time-consuming and error-prone
- **Solution**: Combine smart tools with even smarter humans

INERA
eXtyles | edifix

# Questions?

cgebhard@inera.com