XSL-FO の実践

XSL-FO in practice

Tony Graham
Mentea
13 Kelly's Bay Beach
Skerries, Co Dublin, Ireland
info@mentea.net
@MenteaXML
http://www.mentea.net/

8 July 2014

© 2014 Mentea
All rights reserved.

XSL-FO の実践

XSL-FO in practice

Where Is XSL-FO Used? *5*
Journal Publishing: JATS and PLOS ONE *14*
References *38*
Appendix A – **About** *39*

© 2014 Mentea

© 2014 Mentea

XSL-FO の実践 1

- Introduction
- Where is XSL-FO used?
- Case study: JATS and PLOS ONE

おはようございます。私はグラハムトニーです。どうぞよろしくおねがいいたします。

Who am I? 2

- XML, XSLT, and XSL-FO consulting
- Print and Page Layout Community Group Chair
- W3C Invited Expert
  - XSL 1.0 CR to Recommendation
  - After XSL 1.1 to end
- Japanese Language Task Force (English side)
- Previously ran XSL-List and DSSSList
- Antenna House Alliance Partner
- Lived in Japan 1991–1995

(CR = Candidate Recommendation, which is two stages before becoming a Recommendation)

Where Is XSL-FO Used? 3

My experience includes:

- Conference proceedings
- Automotive maintenance information
- Aircraft manuals
- General publications
- Journals

But no one person has the full picture...

© 2014 Mentea

What Is The Shape Of XSL-FO Today? 4

"Blind monks examining an elephant", an ukiyo-e print by Hanabusa Itchō.

In the Buddhist version of this parable, the blind men come to blows when they can't agree, but I hope we're not going to do that today.

Two World Views 5

- "***CSS clearly and decisively won on the web***, and XSL-FO is being shuttered as a W3C technology, with us absorbing Liam for his expertise in printing tech (which XSL-FO was always somewhat better at) so we can bring CSS up to rough feature parity"
- "Just today I visited my car dealership to pick up my car, my propane tank was filled, I had a doctor's appointment and FedEx delivered a few packages. ***Everyone of them resulted in pages of documentation in print. Much of it probably from XSL FO sources*** (I happen to know two of them were definitely)."

But while, in late 2012, a member of the CSS WG talked about the end of XSL-FO on the web, in late 2013, a representative of a formatter vendor was upbeat about the number of places that XSL-FO is used off the web.

© 2014 Mentea

Announcements, 2013 – www-xsl-fo@w3.org 6

- February – XSLFast

www-xsl-fo@w3.org is the W3C mailing list specific to XSL-FO, and the introduction of the spec directs people there, so the single announcement in 2013 on www-xsl-fo@w3.org would indicate that there's not much activity in XSL-FO land.

Announcements, 2013 – vendor websites 7

- February – Ecrion
- April – Compart
- June – **Antenna House**
- July – Ecrion
- August – Ecrion, RenderX
- September – **Antenna House**
- October – **Antenna House**, Compart, Altova
- November – **Antenna House**
- December – Ecrion, **Antenna House**

However, away from the developer's mailing list, multiple vendors made multiple product announcements on their own websites during 2013.

The announcements indicate activity, but at best the activity is visible only to each vendor's customers and to those who actively look for it.

Print & Page Layout Community Group Survey 8

- Introducing the Community Group
- Survey and results

Another viewpoint can be had from the profiles of the people who responded to a survey run by the Print & Page Layout Community Group at W3C.

Print & Page Layout Community Group 9

"The Print and Page Layout Community Group is open to all aspects of page layout theory and practice. We can and will cover everything from the Crystal Goblet through to specifications and on to the nitty-gritty of writing stylesheets. You will find XSL-FO discussed here, but you will also find other stylesheet languages, and all are equally welcome."

The current group description on the Print & Page Layout Community Group web page.

© 2014 Mentea

History 10

- Started after XML Prague 2012
- Survey on “What to do next?” in January 2014
- Survey advertised on XSL, XSL-FO, XML, and CSS mailing lists, and on Twitter
- 64 responses
  - Twice as many as W3C XSL-FO requirements survey in 2008
- Look at responses about current usage

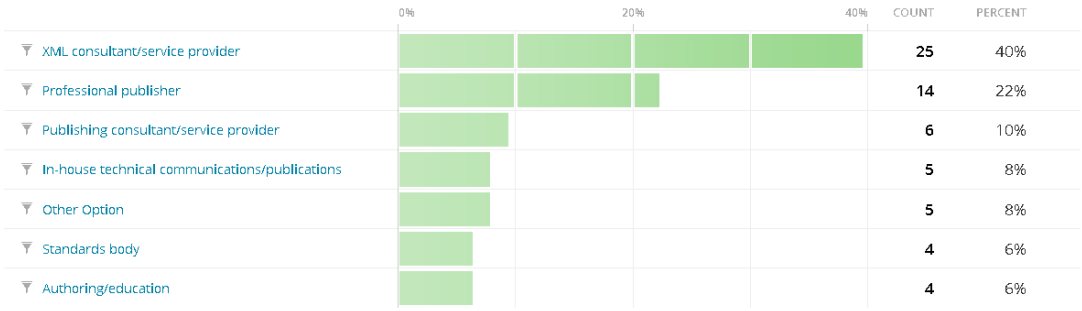
The Print & Page Layout Community Group started out focused on XSL-FO, but re-examined its priorities towards the end of 2013. As part of that, the CG ran a survey on “What to do next?” and advertised it widely on mailing lists and Twitter.

The initial questions in the survey sought information about the person responding, while later questions covered what to do next. The responses to the initial questions, therefore, can give a picture of the XSL-FO usage among those who responded.

Involvements With Markup 11

Question 01 — My involvement is with:

Answers 63 (98%) Skips 1 (2%)



This first question asked what people were involved with.

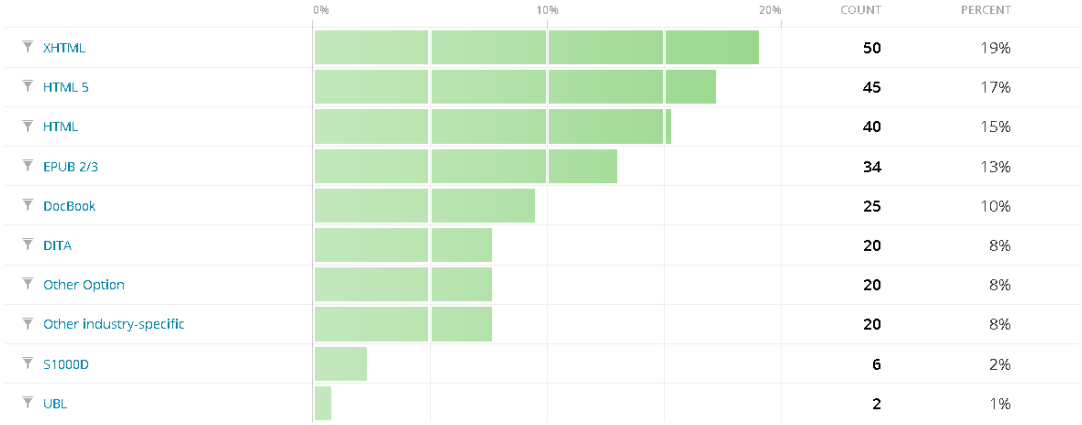
Responding was voluntary, so it’s not surprising that many responders were already involved with XML, though nearly a third were primarily involved with publishing.

© 2014 Mentea

Standards In Use 12

Question 02 — Markup-related standards I use include:

Answers 64 (100%) Skips 0 (0%)



The second question asked what markup-related standards people used. Multiple responses were allowed on this question.

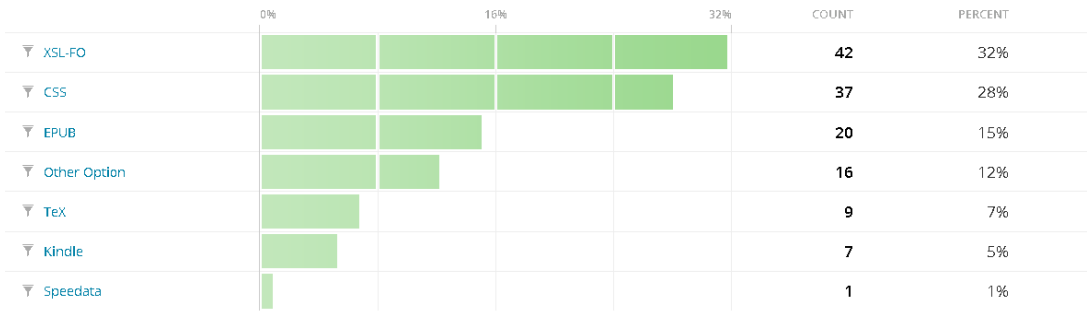
Most people use multiple forms of HTML, and more than half use EPUB.

It's likely also that the people using DocBook, etc., are transforming their XML to make HTML and EPUB.

"Print" Technology 13

Question 03 — I 'print' using:

Answers 64 (100%) Skips 0 (0%)



The third question asked what technologies people are using and provided a mixed selection of technologies to choose from, as well as an "other option" option.

Multiple responses were also allowed on this question, and on average the responders selected two technologies.

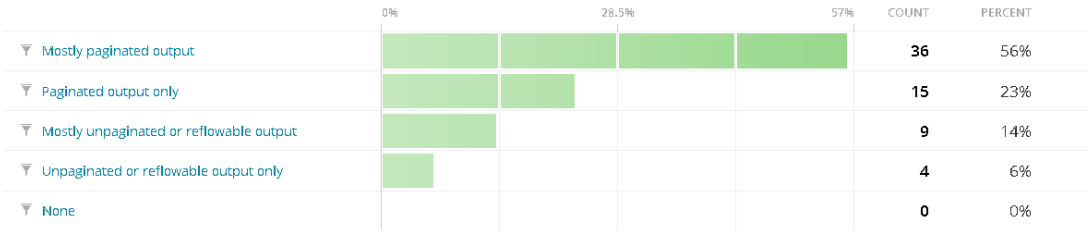
XSL-FO was the most used technology among the responders, which we know were mostly XML professionals and professional publishers. CSS usage wasn't far behind, but while many people must have selected both XSL-FO and CSS, there unfortunately wasn't a question about relative proportions of the usage of both technologies.

© 2014 Mentea

Output Format 14

Question 04 — My 'print' output is:

Answers 64 (100%) Skips 0 (0%)



	COUNT	PERCENT
Mostly paginated output	36	56%
Paginated output only	15	23%
Mostly unpaginated or reflowable output	9	14%
Unpaginated or reflowable output only	4	6%
None	0	0%

The next question asked about the format of the output, with the implication that it was produced using the technologies selected in the previous question.

Although in the previous question more than half of the responders said they use CSS, the responders were mostly (and for nearly a quarter, exclusively) producing paginated output.

Unfortunately, there wasn't a question about which technologies were used for paginated output, so we don't know the balance between XSL-FO usage and CSS usage.

© 2014 Mentea

Interests In Activities 15

Question 05 — My interest in ... is as ...

Answers 64 (100%) Skips 0 (0%)

	SPECIALIST	GENERALIST	NOVICE	INTERESTED BYSTANDER	UNINTERESTED
Page layout	**29**	28	4	0	0
Single source/multiple outputs	**39**	18	4	1	1
XML	**53**	6	0	0	1
HTML/XHTML/HTML 5	**33**	26	5	0	0
CSS	20	**36**	5	1	0
XSL-FO	**23**	17	5	8	6
Digital Publishing	**30**	19	8	2	2
Writing software	**25**	23	4	8	1
Standards development	15	**20**	5	20	1

The final survey question that's useful for today asked people to rate themselves for their level of interest in multiple areas.

This is a "temperature graph" for what people were interested in. Multiple answers were allowed for this question.

From what we know from the first question on involvement, it's not surprising that most people described themself as an XML expert.

The large number of programmers also doesn't reflect the general population, but the number may include people counting XSLT as their programming specialty.

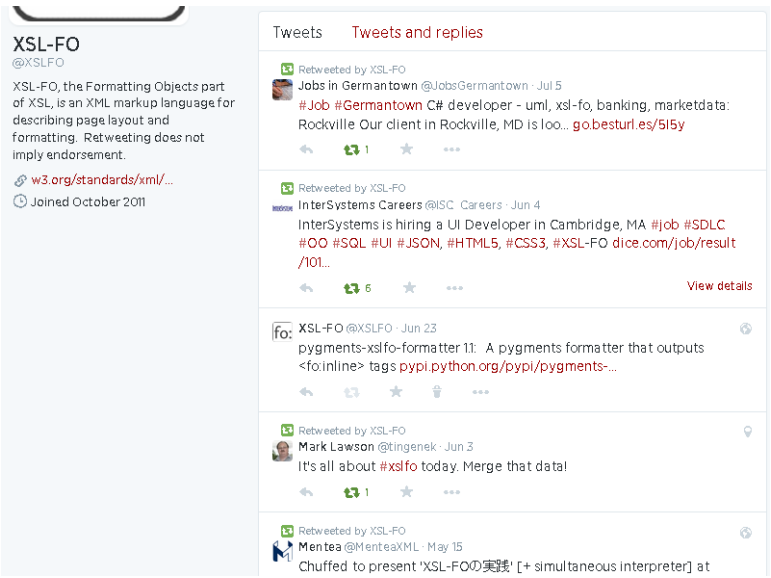
More than half the responders identified themselves as CSS generalists, but there was more of a divide for XSL-FO between specialists and people who were uninterested or a bystander.

Summary: PPL CG Survey 16

- XML professionals and publishers are producing mostly paginated output
- XSL-FO is of interest to people interested in print & page layout
- More people use it than consider themselves experts in it

© 2014 Mentea

XSL-FO On Twitter 17



The `@XSLFO` Twitter account retweets mentions of XSL-FO.

There's usually one job posting a month or so.

Number of followers growing steadily but still not large

DITA and XSL-FO 18

"I think that the DITA-OT community is the largest user group for FOP. If you look at LinkedIn, the DITA awareness group has 552,722 members, the DITA for Small Teams has 64,551 members, DITA Metrics has 74,388. There are country-specific DITA groups with 15-20,000 members. The Tools for Change for Publishing has 3,175,059 members. DITA Machine Industry for technical docs for automotive and and machinery has 18,463 members.
DITA-OT including FOP has been downloaded 344 times this week and over 4,100 times in the last 12 months."

I don't know if we're up to the tail of the elephant or the ears, but the next three slides are extracts of recent email messages discussing XSL-FO usage.

This first one is from a message to the `fop-dev` mailing list in March 2014 from Ron Wheeler, a member of the DITA-OT community.

Aircraft Maintenance and XSL-FO 19

"For my previous employer I was producing 30,000 pages a month with less than $2000 worth of software."

This is from a message from Michael Hahn to the Print & Page Layout Community Group mailing list in December 2013 referring to his previous job in Technical Publications for a light-aircraft manufacturer where he used XSL-FO to produce thousands of pages per month.

© 2014 Mentea

Publishing and XSL-FO 20

“XSL-FO is very, very widely used in professional publishing; the largest publishers use XSL-FO (and in some cases XHTML + CSS) for most of their fiction and mainstream texts.
There are relatively few publishers using it, but between them they produce most of the books you see - although there's much more diversity in Europe than in the USA in that regard.”

This last one is from Liam Quin of W3C to the `XSL-List` mailing list in May 2014 on a thread discussing XSL-FO usage for professional publishing.

Summary: Shape Of XSL-FO Today? 21

- No-one has the complete picture
- Many use its output without knowing
- More run XSL-FO than are experts

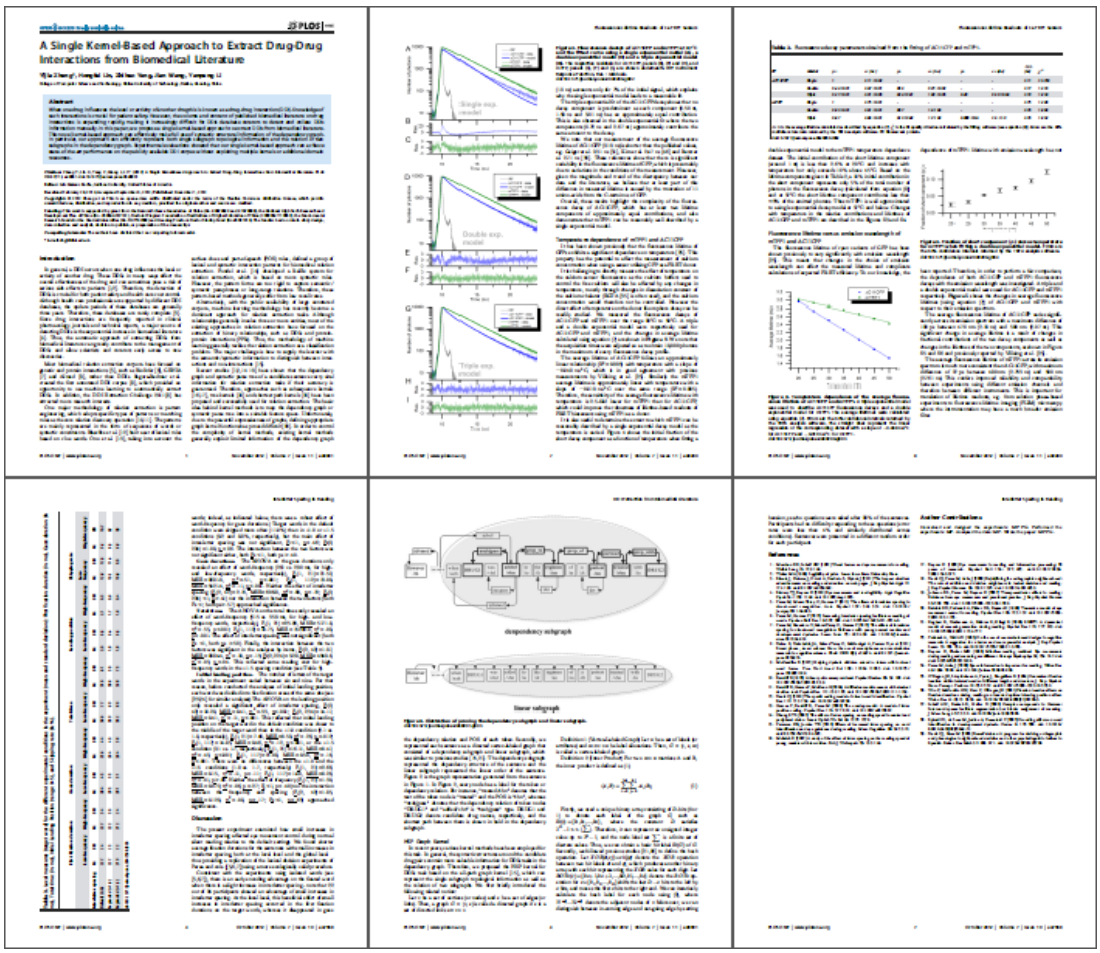
We have seen a variety of views. Information about XSL-FO usage isn’t available all in one place, and information such as new product announcements isn’t widely disseminated. We saw from the Print & Page Layout CG survey that while some people are XSL-FO experts, there are others who aren’t, so that while there are many downloads of the DITA Open Toolkit, many more will use the XSL-FO processor in the toolkit than will modify the XSL-FO. Similarly, there are many more people using the books or maintenance manuals produced using XSL-FO than use an XSL-FO processor.

Questions? 22

質問がありますか。

© 2014 Mentea

Journal Publishing: JATS and PLOS ONE



In this half of the presentation, I will talk in general about journal publishing from XML using JATS and then talk specifically about work done formatting articles for the PLOS ONE journal.

© 2014 Mentea

What is JATS? 24

`http://jats.nlm.nih.gov/`

- Journal Article Tag Suite
- For STM journal articles
  - Related “BITS” schema for STM books
- Used by journal publishers and archives around the world
- Successor to “National Library of Medicine (NLM) Journal Archiving and Interchange Tag Suite”

The NLM Tag Suite project was started by the National Center for Biotechnology Information (NCBI) at NLM in 2002 as the schema for articles submitted to PubMed. When usage of the NLM Tag Suite grew to be more than just for submitting to PubMed, the Tag Suite moved to the NISO standards organisation, and the name changed to “Journal Article Tag Suite” to better reflect the wider usage.

(“STM” = “Scientific, Technical, and Medical”)

Who Uses JATS? 25

- Over 2 million XML articles in PubMed Central (in 2011)
- Nature
  - Announced two Open Access journals in January 2014
- International Standards Organisation (ISO)
  - ISO Standards Tags Set (ISO STS) is an adaption of JATS
- Optical Society of America
  - Converted 95 years of legacy journal content to XML
- Substance Abuse and Mental Health Services Administration (SAMHSA), USA
  - Source for printed and mobile versions of publications
- 学術情報 XML 推進協議会
  - XML Scholarly Publishing Association (XSPA)
- Public Library of Science (PLOS), USA
  - Open Access journals

© 2014 Mentea

What Colour is Your JATS? 26

Three models for different stages in life of an article:

- **Article Authoring Tag Set**
  - For authoring well-structured articles
  - Very prescriptive, no generated content
- **Journal Publishing Tag Set**
  - What you'd accept for publishing
  - Restrictive, limited options
- **Journal Archive and Interchange Tag Set**
  - The output of a published article or where you have no control over content
  - Most permissive

The JATS Tag Suite defines three "tag sets" that you can use as-is or use as the base for your own customisation of JATS. The tags sets each cover a different stage in the life of an article, and they are commonly referred to by their identifying colour: orange, blue, or green.

(The difference between "tag suite" and "tag set" isn't all that obvious even to English speakers.)

`address` in Article Authoring Tag Set 27

```
((institution | institution-wrap)*, addr-line*, country, phone*, fax*, email*, ext-link*, uri*)
```

- Fixed order of elements
- Requires one `country`

We'll look now at the `address` element to illustrate the differences between the tag sets.

The Orange tag set is designed to enable structured authoring.

`address` in the Article Authoring Tag Set requires its elements in a certain order and requires that `country` is always part of the address.

© 2014 Mentea

address in Journal Publishing Tag Set 28

```
(addr-line | country | fax | institution | institution-wrap | phone |
email | ext-link | uri)*
```

- Same elements allowed
- No fixed order
- Zero or multiple of everything okay
- Allows multiple country

In the Journal Publishing Tag Set, the same elements are allowed, but there's no restrictions on order or occurrence of the elements. The Blue tag set is more what you'd accept for a publishing process where you have less control over the content to be published. The blue highlights in the content model show the differences from the Orange tag set.

address in Journal Archive and Interchange Tag Set 29

```
(addr-line | country | fax | institution | institution-wrap | phone |
email | ext-link | uri | label | x)*
```

- Same elements in any order and any number
- Additional elements:
  - label – "Label (of an Equation, Figure, Reference, etc.)"
  - x – "X - Generated Text and Punctuation"

The Journal Archive and Interchange Tag Set is just as relaxed, and it also allows label and x for representing generated content – such as figure numbers, etc. – that was added as part of the publishing process. The green highlights in the content model show the extra elements allowed in the Green tag set.

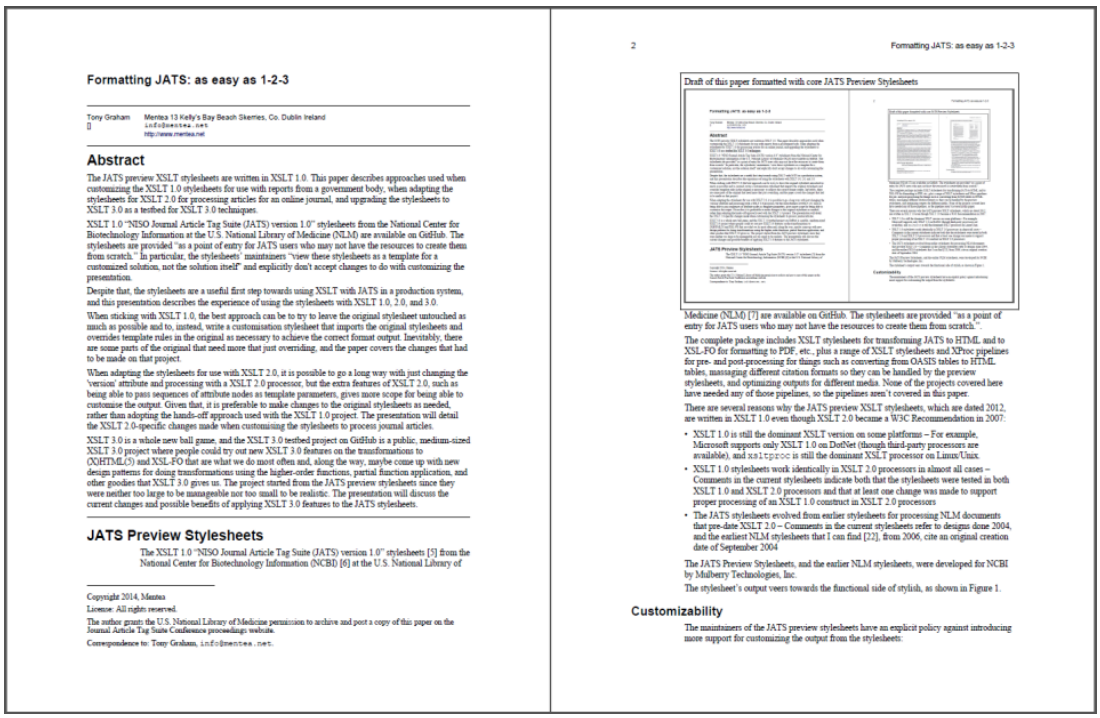
JATS Preview Stylesheets 30

https://github.com/NCBITools/JATSPreviewStylesheets

- XSLT 1.0
- Public domain
- No copyright issues
- Developed for NCBI by Mulberry Technologies

© 2014 Mentea

JATS Preview Stylesheets Output 31



Formatting JATS: as easy as 1-2-3

Tony Graham Mentea 13 Kelly's Bay Beach Skerries, Co. Dublin Ireland info@mentea.net http://www.mentea.net

Abstract

The JATS preview XSLT stylesheets are written in XSLT 1.0. This paper describes approaches used when customizing the XSLT 1.0 stylesheets for use with reports from a government body, when adapting the stylesheets for XSLT 2.0 for processing articles for an online journal, and upgrading the stylesheets to XSLT 3.0 as a testbed for XSLT 3.0 techniques.

XSLT 1.0 "NISO Journal Article Tag Suite (JATS) version 1.0" stylesheets from the National Center for Biotechnology Information at the U.S. National Library of Medicine (NLM) are available on GitHub. The stylesheets are provided "as a point of entry for JATS users who may not have the resources to create them from scratch." In particular, the stylesheets' maintainers "view these stylesheets as a template for a customized solution, not the solution itself" and explicitly don't accept changes to do with customizing the presentation.

Despite that, the stylesheets are a useful first step towards using XSLT with JATS in a production system, and this presentation describes the experience of using the stylesheets with XSLT 1.0, 2.0, and 3.0.

When sticking with XSLT 1.0, the best approach can be to try to leave the original stylesheet untouched as much as possible and to, instead, write a customisation stylesheet that imports the original stylesheets and overrides template rules in the original as necessary to achieve the correct format output. Inevitably, there are some parts of the original that need more that just overriding, and the paper covers the changes that had to be made on that project.

When adapting the stylesheets for use with XSLT 2.0, it is possible to go a long way with just changing the 'version' attribute and processing with a XSLT 2.0 processor, but the extra features of XSLT 2.0, such as being able to pass sequences of attribute nodes as template parameters, gives more scope for being able to customise the output. Given that, it is preferable to make changes to the original stylesheets as needed, rather than adopting the hands-off approach used with the XSLT 1.0 project. The presentation will detail the XSLT 2.0-specific changes made when customising the stylesheets to process journal articles.

XSLT 3.0 is a whole new ball game, and the XSLT 3.0 testbed project on GitHub is a public, medium-sized XSLT 3.0 project where people could try out new XSLT 3.0 features on the transformations to (X)HTML(5) and XSL-FO that are what we do most often and, along the way, maybe come up with new design patterns for doing transformations using the higher-order functions, partial function application, and other goodies that XSLT 3.0 gives us. The project started from the JATS preview stylesheets since they were neither too large to be manageable nor too small to be realistic. The presentation will discuss the current changes and possible benefits of applying XSLT 3.0 features to the JATS stylesheets.

JATS Preview Stylesheets

The XSLT 1.0 "NISO Journal Article Tag Suite (JATS) version 1.0" stylesheets [5] from the National Center for Biotechnology Information (NCBI) [6] at the U.S. National Library of

Copyright 2014, Mentea
License: All rights reserved.
The author grants the U.S. National Library of Medicine permission to archive and post a copy of this paper on the Journal Article Tag Suite Conference proceedings website.

Correspondence to: Tony Graham, info@mentea.net.

Draft of this paper formatted with core JATS Preview Stylesheets

Medicine (NLM) [7] are available on GitHub. The stylesheets are provided "as a point of entry for JATS users who may not have the resources to create them from scratch.".

The complete package includes XSLT stylesheets for transforming JATS to HTML and to XSL-FO for formatting to PDF, etc., plus a range of XSLT stylesheets and XProc pipelines for pre- and post-processing for things such as converting from OASIS tables to HTML tables, massaging different citation formats so they can be handled by the preview stylesheets, and optimizing outputs for different media. None of the projects covered here have needed any of those pipelines, so the pipelines aren't covered in this paper.

There are several reasons why the JATS preview XSLT stylesheets, which are dated 2012, are written in XSLT 1.0 even though XSLT 2.0 became a W3C Recommendation in 2007:

- XSLT 1.0 is still the dominant XSLT version on some platforms – For example, Microsoft supports only XSLT 1.0 on DotNet (though third-party processors are available), and `xsltproc` is still the dominant XSLT processor on Linux/Unix
- XSLT 1.0 stylesheets work identically in XSLT 2.0 processors in almost all cases – Comments in the current stylesheets indicate both that the stylesheets were tested in both XSLT 1.0 and XSLT 2.0 processors and that at least one change was made to support proper processing of an XSLT 1.0 construct in XSLT 2.0 processors
- The JATS stylesheets evolved from earlier stylesheets for processing NLM documents that pre-date XSLT 2.0 – Comments in the current stylesheets refer to designs done 2004, and the earliest NLM stylesheets that I can find [22], from 2006, cite an original creation date of September 2004

The JATS Preview Stylesheets, and the earlier NLM stylesheets, were developed for NCBI by Mulberry Technologies, Inc.

The stylesheet's output veers towards the functional side of stylish, as shown in Figure 1.

Customizability

The maintainers of the JATS preview stylesheets have an explicit policy against introducing more support for customizing the output from the stylesheets:

JATS-Con paper formatted with default JATS Preview stylesheets

What is Provided? 32

- Preprocessing
  - Convert OASIS tables to HTML tables
  - Massage citation format
  - Some require XSLT 2.0
- Formatting
  - XML to HTML
  - **XML to XSL-FO for formatting as PDF**
- Post-processing
  - HTML to XHTML for MathML

The only part that I've needed to use, and the only part being covered, is the transformation to XSL-FO and formatting to PDF.

© 2014 Mentea

Customizability 33

“These stylesheets are **provided as a point of entry for JATS users** who may not have the resources to create them from scratch. Because there are many varied implementations of JATS, you should have no expectation that these stylesheets will create production ready files in any arbitrary system. Instead, the stylesheets should be customized for your particular needs.”

“Because **we view these stylesheets as a template for a customized solution, not the solution itself**, we will accept changes that fix an actual bug, but we will not merge in changes that we view as “customization”. For example, we will accept changes that fix a problem which otherwise leads to failure in creating a final output file, but we will not accept changes that focus on presentational aspects of the final output (such as font changes, margin changes, graphics sizing, etc).”

Statement about customisation from JATSPreviewStylesheets README with added emphasis.

XSLT features supporting customizability 34

- Templates
- Modular stylesheets
- Named attribute sets

Templates 35

- `match` matches a context in source XML
- Content of `xsl:template` instantiated when template is applied

```
<xsl:template match="td">
  <fo:table-cell xsl:use-attribute-sets="td">
    <xsl:call-template name="process-table-cell"/>
  </fo:table-cell>
</xsl:template>
```

Elements in the body of the template not in the XSLT namespace are copied to the result, and elements and attributes in the XSLT namespace are acted on by the XSLT processor.

© 2014 Mentea

Modular stylesheets 36

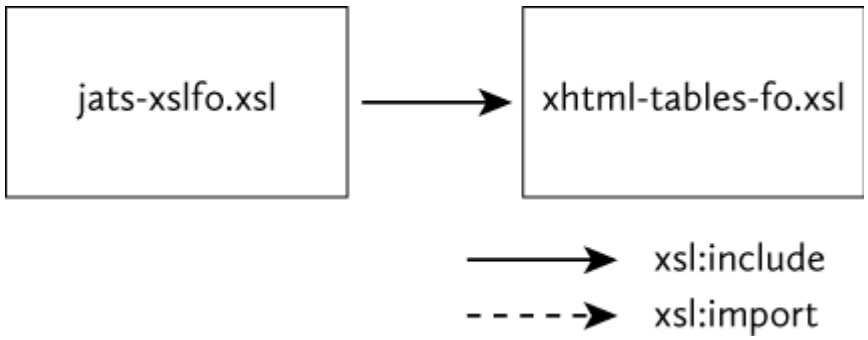
```
<xsl:include
  href = uri-reference />
```

- `href` refers to other stylesheet
- Children of other `xsl:stylesheet` replace `xsl:include`

```
<xsl:import
  href = uri-reference />
```

- `href` refers to other stylesheet
- Imported definitions and template rules *not* part of importing stylesheet
- Have lower *import precedence*

Imports in JATS XSL-FO preview 37



There a more interesting block diagram later.

Overriding templates 38

- Template in importing stylesheet overrides same context in imported
- Good when overriding complete function of template
- Extra overhead if you just want to change one little thing

© 2014 Mentea

Attribute sets 39

- Named set of attribute definitions
- Use in multiple places
- Definitions evaluated in each context where used

```
<xsl:attribute-set name="fig">
  <xsl:attribute name="keep-together.within-page"
      >always</xsl:attribute>
  <xsl:attribute name="id">
    <xsl:value-of select="generate-id()" />
  </xsl:attribute>
</xsl:attribute-set>
```

Since attribute definitions in attribute sets are evaluated each time the attribute set is used, the value of the `id` attribute will be unique to each context.

JATS Preview supporting customizability 40

- Global variables
- Attribute sets
- Named templates

© 2014 Mentea

Example customization 41

- Add to attribute set from JATS stylesheets

```
<xsl:attribute-set name="td">
  <xsl:attribute name="line-stacking-strategy"
      >max-height</xsl:attribute>
</xsl:attribute-set>
```

- New attribute set reusing merged `td` attribute set

```
<xsl:attribute-set name="td-small"
    use-attribute-sets="td">
  <xsl:attribute name="line-height">10pt</xsl:attribute>
  <xsl:attribute name="border">none</xsl:attribute>
  <xsl:attribute name="padding-top">0pt</xsl:attribute>
  <xsl:attribute name="padding-bottom">0pt</xsl:attribute>
</xsl:attribute-set>
```

- Override JATS stylesheet in more-specific context

```
<xsl:template
    match="td[ancestor::table[@style = 'small']]">
  <fo:table-cell xsl:use-attribute-sets="td-small">
    <xsl:call-template name="process-table-cell"/>
  </fo:table-cell>
</xsl:template>
```

The `xsl:attribute-set` extends the 'td' defined in the JATS Preview stylesheet.

The new 'td-small' attribute set includes the attribute definitions from all declarations for the 'td' attribute set plus the definitions contained in its definition.

The template matches on a more-specific context than the general-purpose template for `td` in the JATS Preview stylesheets, so in those particular contexts, the XSLT processor uses this template, which adds a different set of attributes to the generated `fo:table-cell` but which still uses the 'process-table-cell' named template from the JATS Preview stylesheets as is used in the original template for `td`.

This illustrates in a nutshell how a customisation is able to extend, override, and reuse the constructs in the core JATS Preview stylesheets.

Summary: JATS Preview 42

- XSLT 1.0
- Not accepting customisations into core
- Stylesheet structure facilitates customisations

Summary: JATS 43

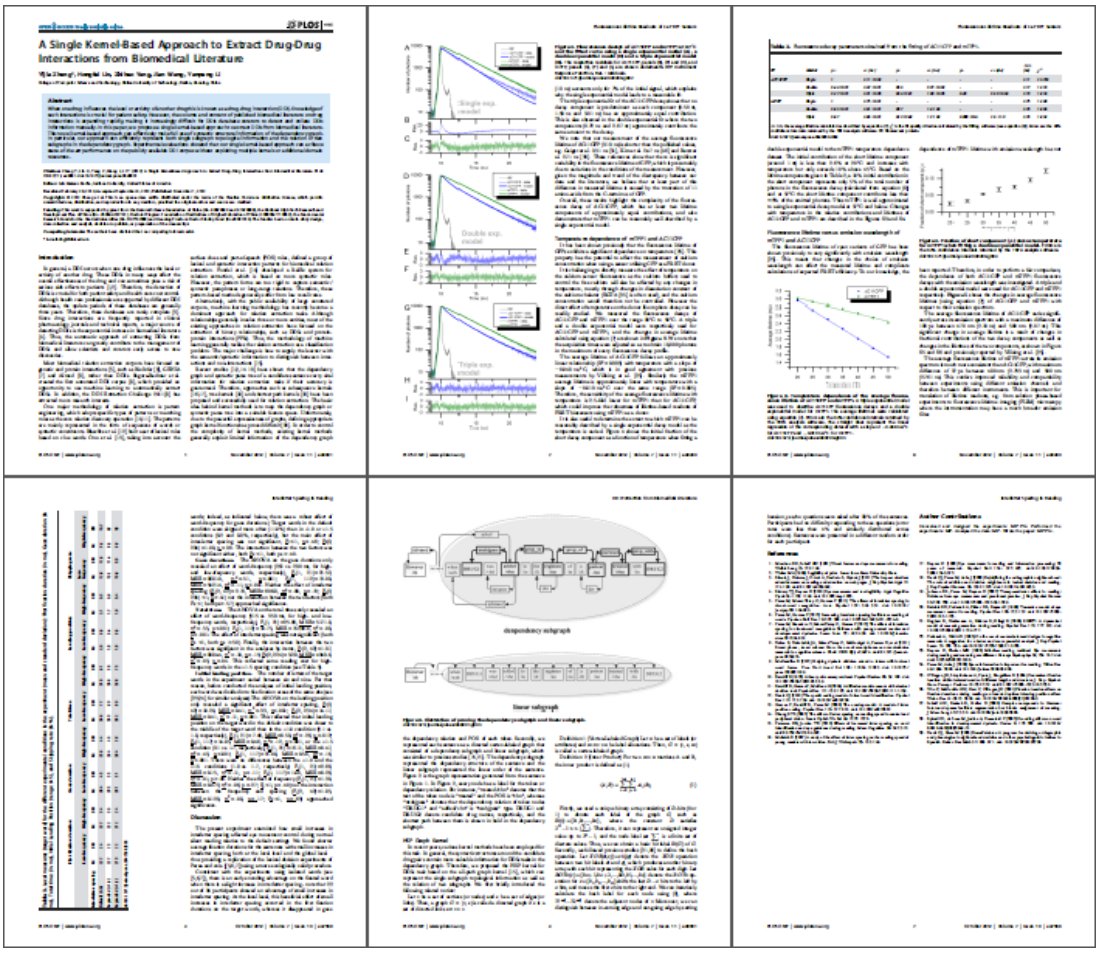
- Three tags sets for different stages
- JATS Preview stylesheets:
  - Explicitly don't support customisation
  - Good basis for your own customization
- Customise by:
  - Layer on top of existing styleheets
  - Modify your copy of the stylesheets

© 2014 Mentea

PLOS ONE 44

- Peer-reviewed, open-access, online publication
- Public Library of Science
- JATS/NLM markup
- Lights-out batch formatting with XSL-FO
- Previously produced use 3B2 and (presumably) manual fix-up
- XSLT 2.0 because…
  - Big differences in metadata, figure, table handling
  - Needed vendor extensions
- Customisation…
  - Modified version of `jats-xslfo.xsl`
  - Additional XSLT modules

PLOS ONE Samples 45



PLOS ONE "Features" 46

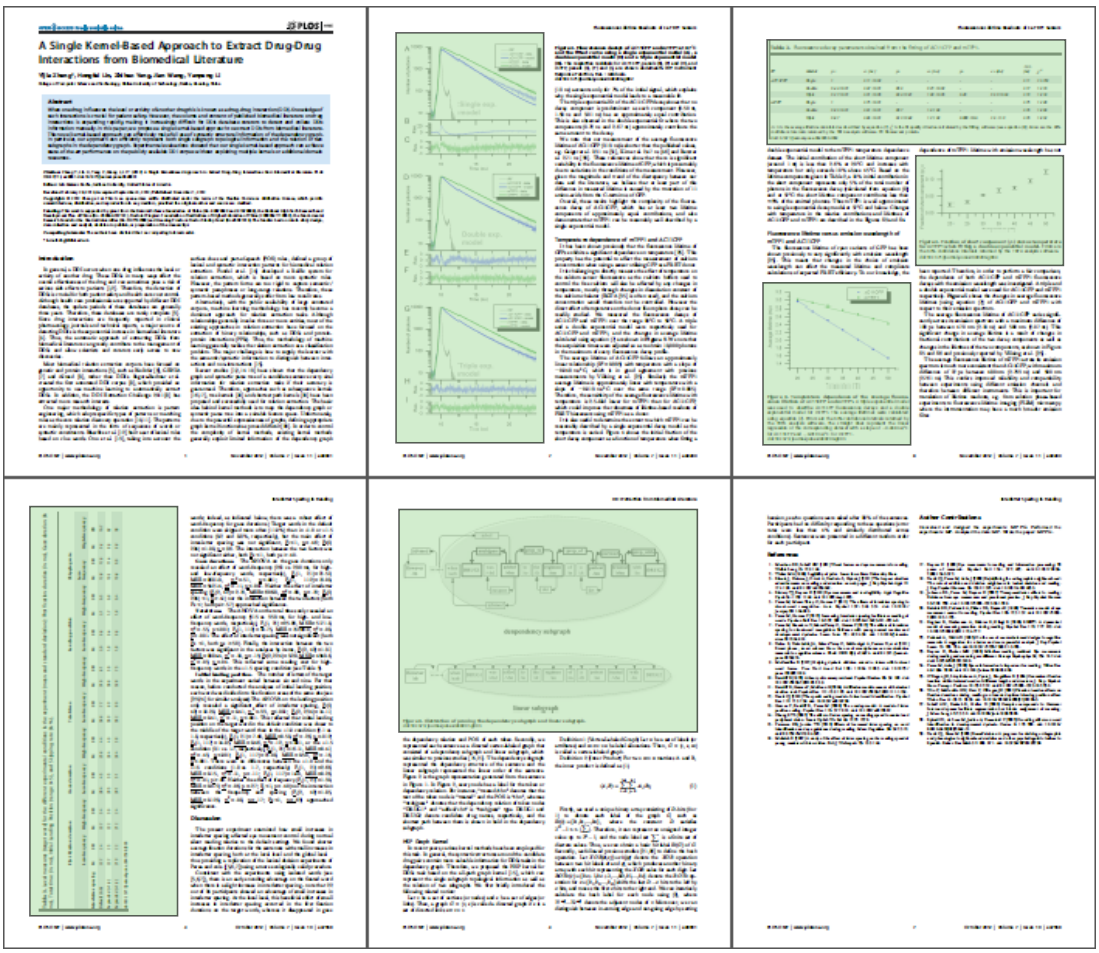
- Figures and tables float to top (or bottom) of page
- Figures column-wide or page-wide
  - No size information in XML
- Figure graphic+caption can't overflow page
- Tables column-wide, page-wide, or page-high
  - Page-high may be single column
  - May be multiple pages
  - No width indication in XML
  - No row spanning (thank goodness!)
- No figures or tables allowed after start of back matter

XSLT/XSL-FO "Features" 47

- Page-wide floats
  - Vendor extension for column-wide
- Floats don't break
- Floats only at top of page
  - Bottom-float extension available but unused
- Graphic size not available to XSLT
- *Fire-and-forget* processing

© 2014 Mentea

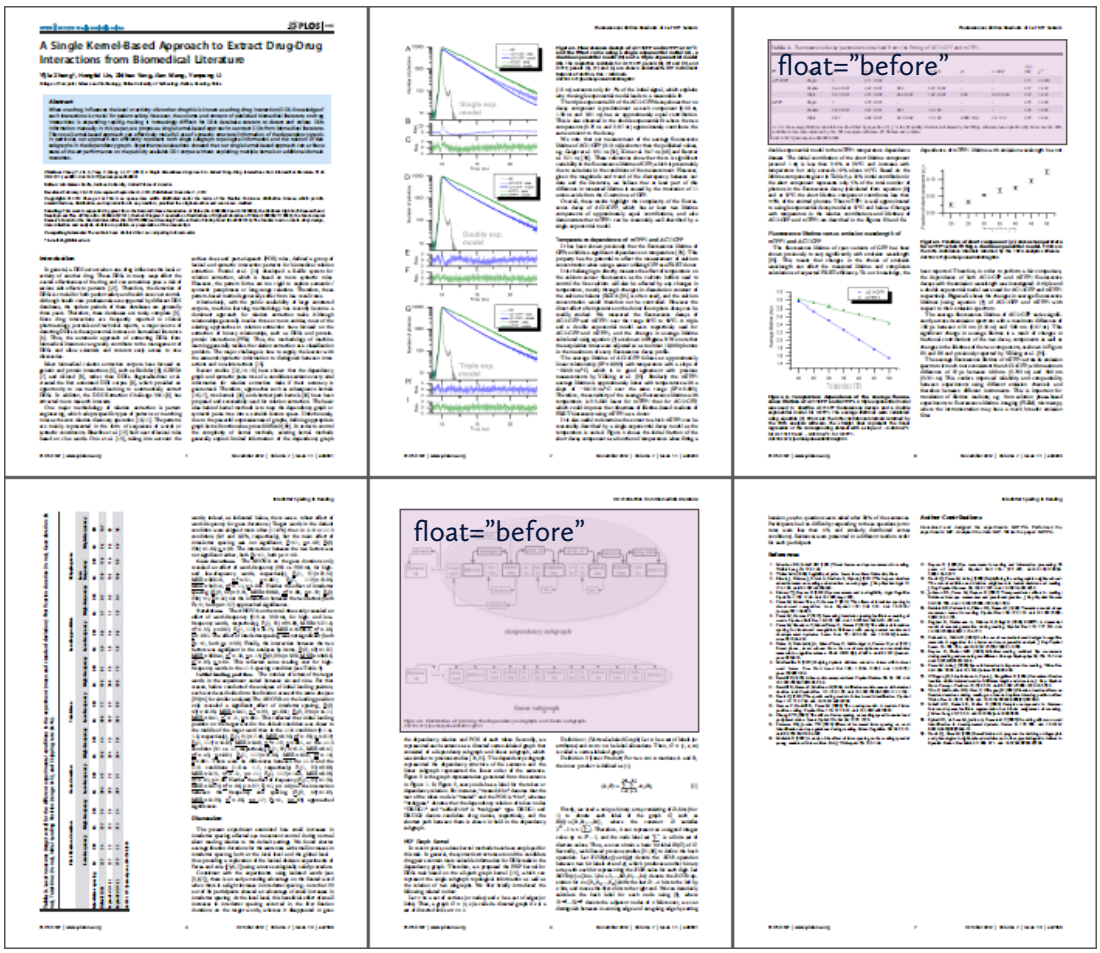
Floats in Sample PLOS Pages



The green areas show floats in the sample PLOS ONE pages.

© 2014 Mentea

Using XSL 1.1 before-float-reference-area

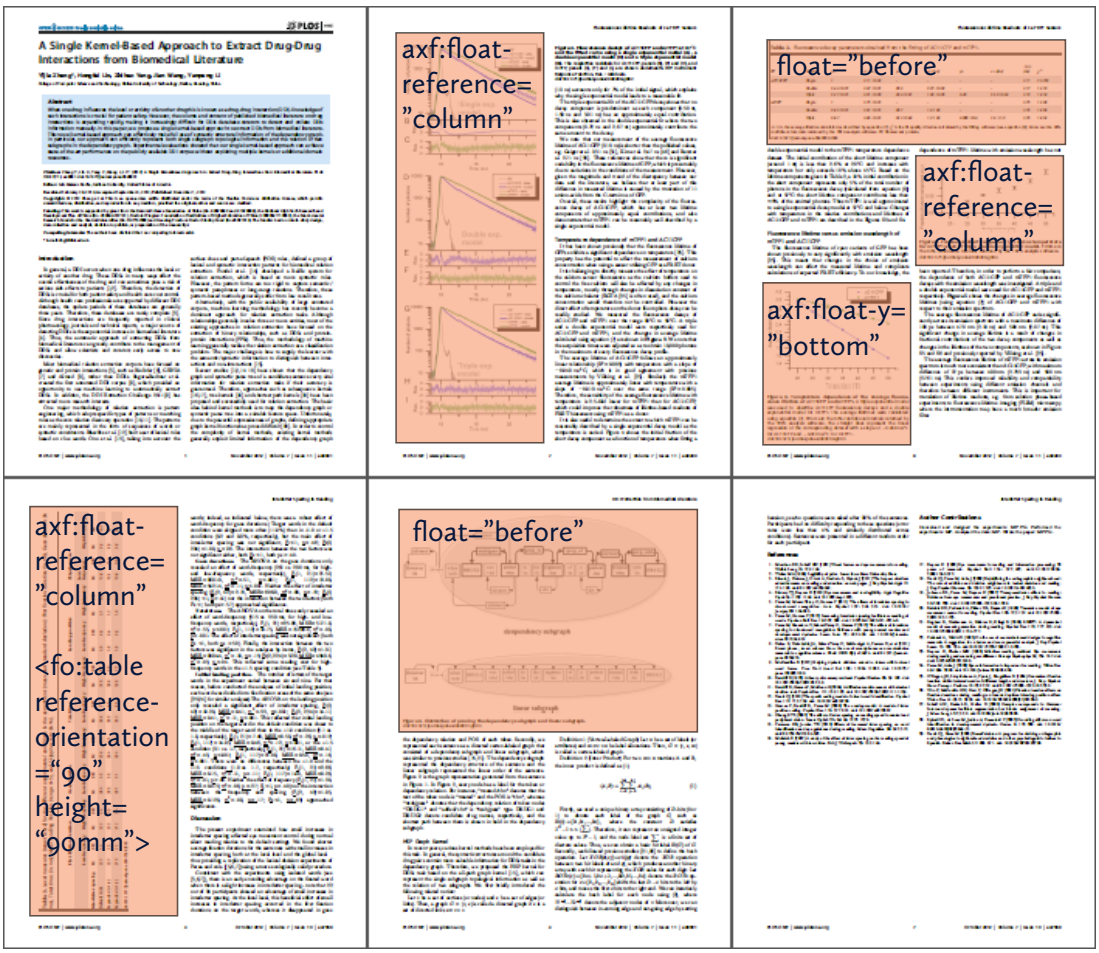


The pink areas highlight the floats that can be represented with the 'before-float-reference-area' in XSL-FO 1.1.

© 2014 Mentea

Using Antenna House Float Extensions

50



The orange areas highlight the floats in PLOS ONE that can be represented using the float extensions in Antenna House Formatter. The text overlaid on the areas illustrate the extension properties and their values.

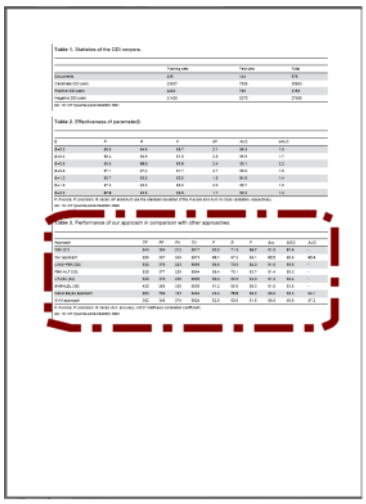
Table Handling

51

- “Pre-format” tables in three widths on *long* pages
  - Column-wide, page-wide, (width of) page-high
  - Prefix table IDs with string indicating width
- Format to area tree XML
- Compare area trees for each table
- Use width with least area and no overflow
- Recreate as multiple `fo:float` if overflows page
  - Re-use table column widths from area tree to remain consistent

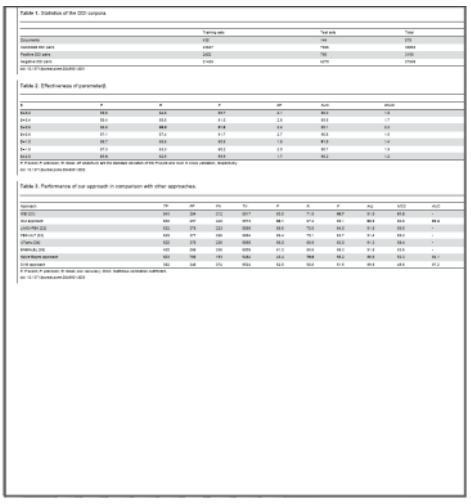
© 2014 Mentea

Picking “Best” Tables



Page-wide

Column-wide



Page-high

Three tables formatted in each of three widths, with preferred versions highlighted.

Sized and Placed Tables

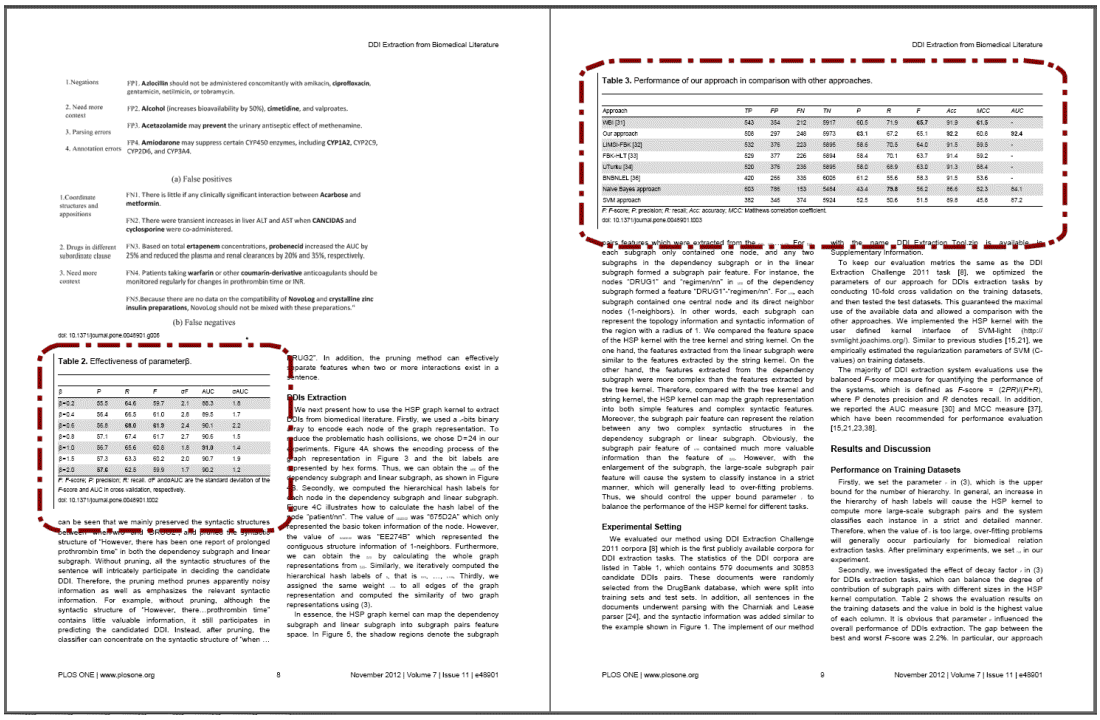


Table 2. Effectiveness of parameter β.

β	P	R	F	dF	AUC	dAUC
β=0.2	55.5	64.6	59.7	2.1	89.3	1.8
β=0.4	56.4	66.5	61.0	2.6	89.8	1.7
β=0.6	56.8	68.0	61.9	2.4	90.1	2.2
β=0.8	57.1	67.4	61.7	2.7	90.6	1.5
β=1.0	56.7	65.6	60.6	1.8	91.0	1.4
β=1.5	57.3	63.3	60.2	2.0	90.7	1.9
β=2.0	57.6	62.5	59.9	1.7	90.3	1.3

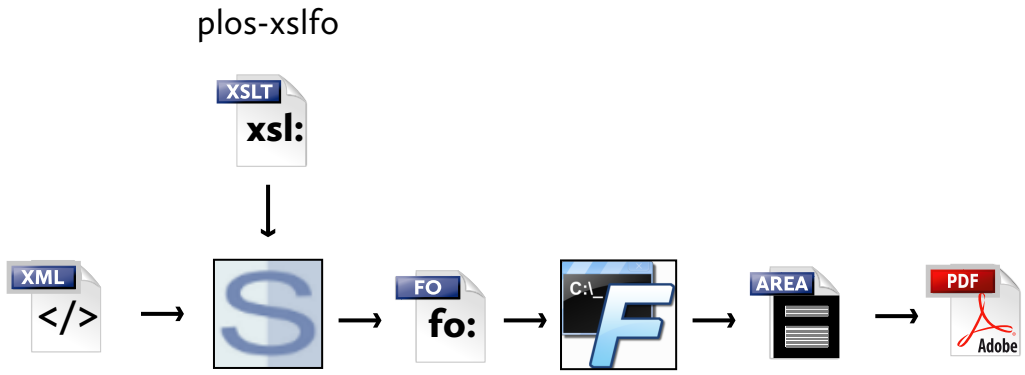
Table 3. Performance of our approach in comparison with other approaches.

Approach	TP	FP	FN	TN	P	R	F	Acc	MCC	AUC
MBI [31]	543	354	212	5917	60.5	71.9	65.7	91.9	61.9	-
Our approach	508	297	246	5973	63.1	67.2	65.1	92.2	60.8	92.4
LIMD-FBK [32]	532	376	223	5895	58.6	70.5	64.0	91.5	59.5	-
FBK-HLT [33]	529	377	226	5894	58.4	70.1	63.7	91.4	59.2	-
UTurku [34]	520	376	235	5895	58.0	68.9	63.0	91.3	58.4	-
ENBNLEL [36]	420	266	335	6005	61.2	55.6	58.3	91.5	53.6	-
Naive Bayes approach	603	795	153	5494	43.4	79.8	56.2	86.6	52.3	84.1
SVM approach	352	346	374	5924	52.5	50.6	51.5	89.8	45.6	87.2

Column-wide and page-wide tables placed on pages.

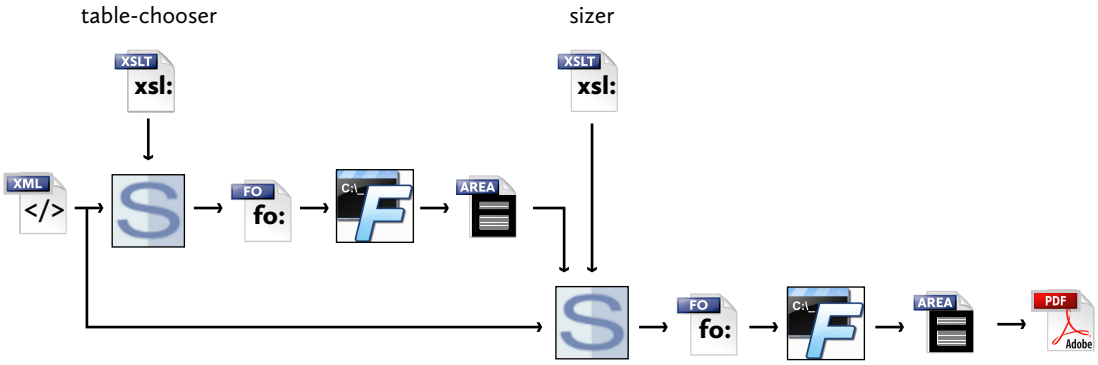
© 2014 Mentea

Usual Processing Model 54



The conventional XSLT–XSL-FO processing model.

Table-handling Processing Model 55



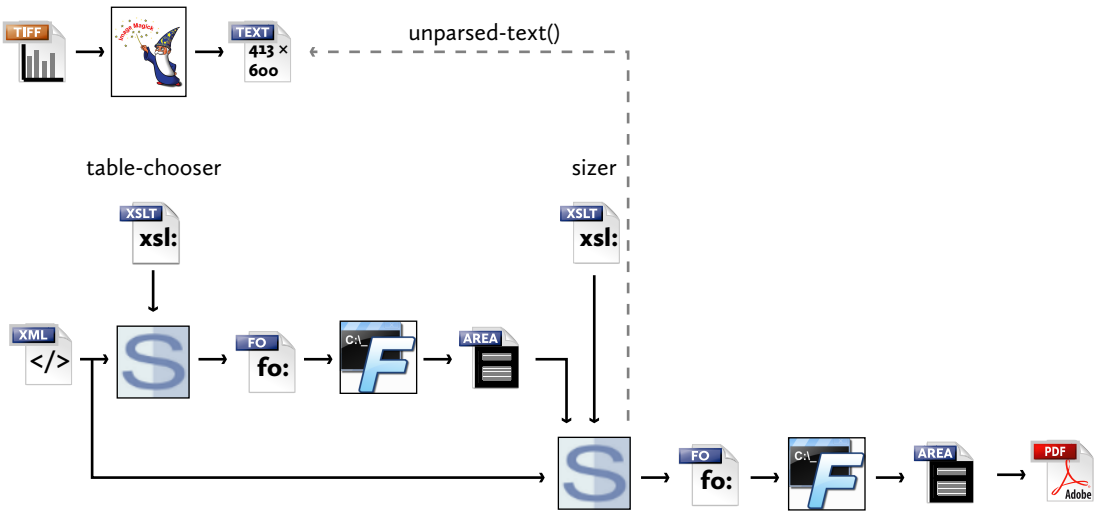
The processing model including preprocessing tables to generate an area tree from which to determine the preferred width for each table.

Graphics Handling 56

- Get TIFF graphics
- ImageMagick `identify` gives graphic size and resolution
- “Pre-format” caption at both widths to get exact size
- Choose best width
- (Possibly) scale down graphic so caption also fits on page

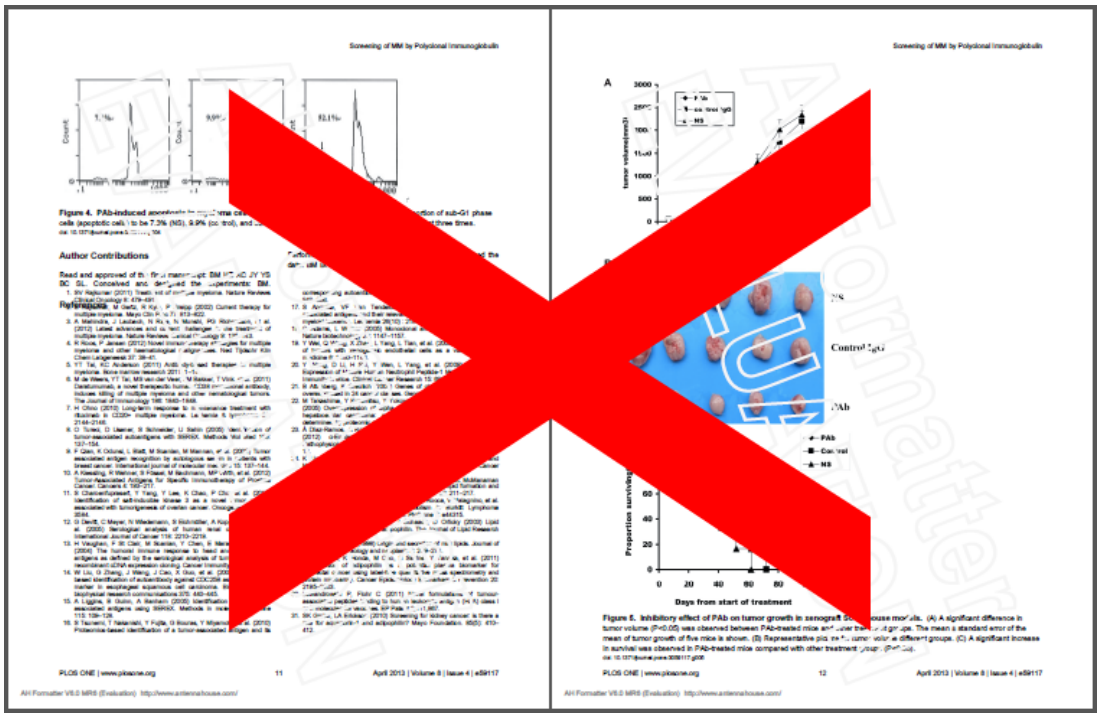
© 2014 Mentea

Figure-handling Processing Model 57



Processing model when graphics handling added.

Floats After Back Matter 58



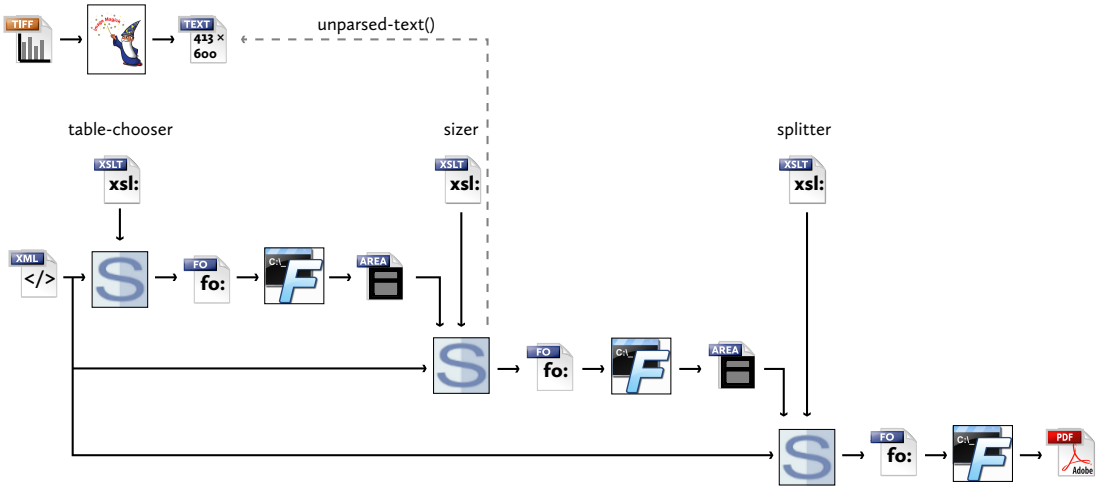
Figures and tables are required to not appear after the start of the back matter.

© 2014 Mentea

Splitting at Back Matter 59

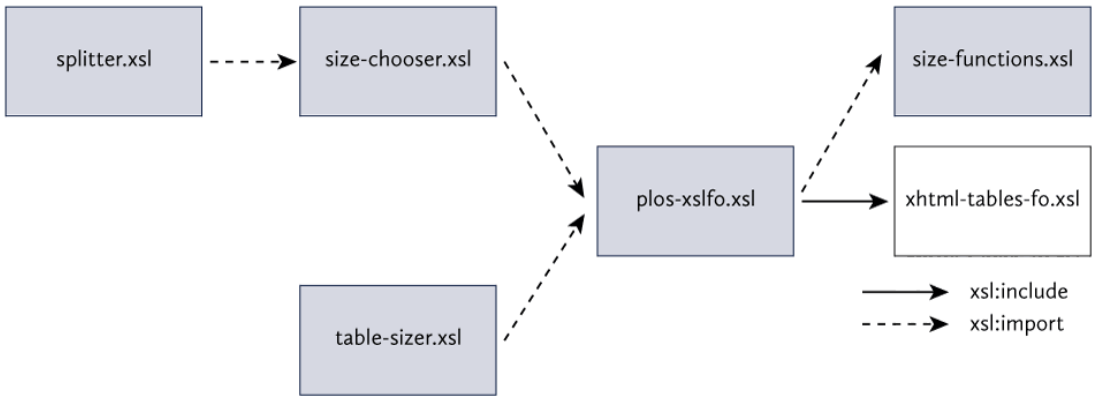
- Format “final” FO with right-width tables and figures to area tree
- Compare positions of first “back” content and last float
  - `back` plus bits from `front`, `body`
- Generate new FO with either one or two `fo:page-seqence`
- If second `fo:sequence`, it contains only back matter so floats in first appear before back matter

Putting It All Together 60



The full processing model.

Import structure 61



All the top-level stylesheets use `plos-xslfo.xsl` for basic formatting.

`splitter.xsl` does everything `size-chooser.xsl` does, and more, so it imports that file rather than importing `plos-xslfo.xsl` directly.

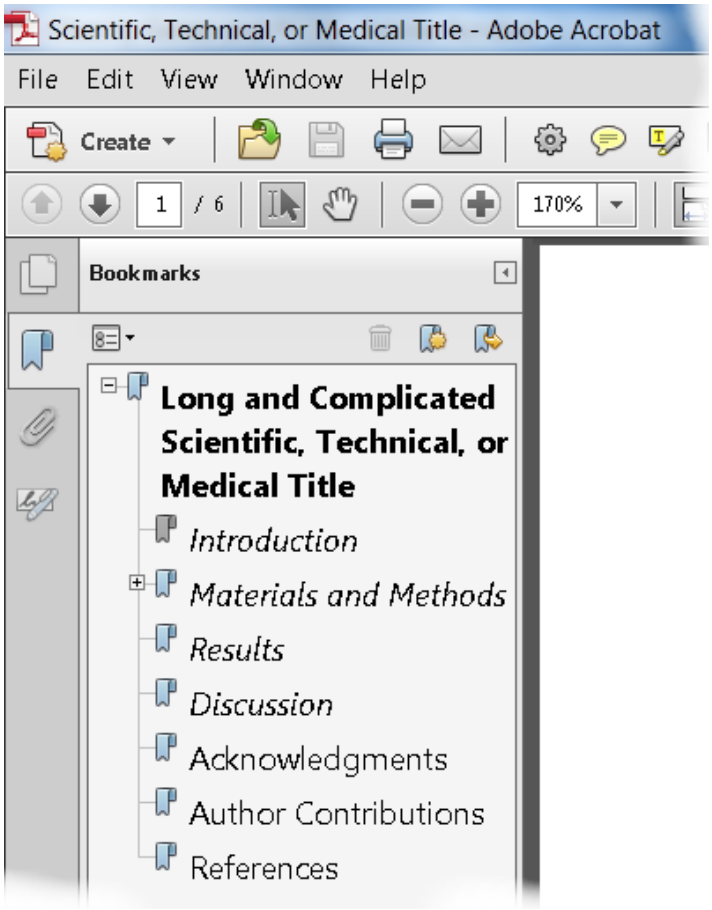
© 2014 Mentea

Other Features 62

- Table of contents and document properties
- Rearranging article metadata

Table of contents and document properties 63

- Bookmarks
  - XSL 1.1
- PDF properties
  - Title, etc. in PDF file
  - Antenna House extension



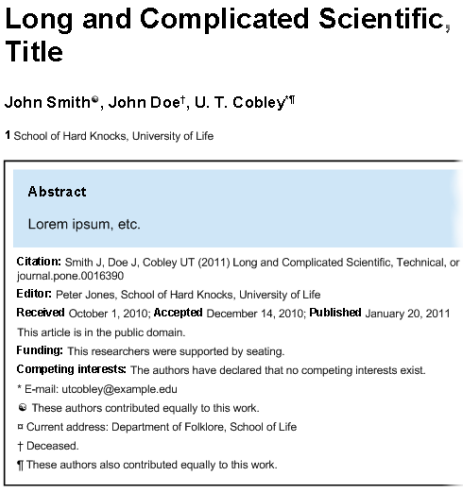
Formatting objects of bookmarks, which can be used to generate bookmarks in PDF, are part of XSL 1.1.

Setting PDF properties, such as the document title that can show in the banner of the PDF viewer, required a vendor extension, however.

© 2014 Mentea

Rearranging Article Metadata 64

```
<article-title>Long and Complicated...
...
<contrib deceased="yes">
  <name>
    <surname>Doe</surname>
    <given-names>John</given-names>
  </name>
  <xref ref-type="aff" rid="aff1">
    <sup>1</sup>
  </xref>
</contrib>
<pub-date pub-type="epub">
  <day>20</day><month>1</month><year>2011</year>
</pub-date>
<volume>6</volume><issue>1</issue>
...
<date date-type="received">
  <day>1</day><month>10</month><year>2010</year>
</date>
<date date-type="accepted">
  <day>14</day><month>12</month><year>2010</year>
</date>
```



Long and Complicated Scientific, Title

John Smith☯, John Doe†, U. T. Cobley¤¶

1 School of Hard Knocks, University of Life

Abstract

Lorem ipsum, etc.

Citation: Smith J, Doe J, Cobley UT (2011) Long and Complicated Scientific, Technical, or journal.pone.0016390

Editor: Peter Jones, School of Hard Knocks, University of Life

Received October 1, 2010; **Accepted** December 14, 2010; **Published** January 20, 2011

This article is in the public domain.

Funding: This researchers were supported by seating.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: utcobley@example.edu

☯ These authors contributed equally to this work.

¤ Current address: Department of Folklore, School of Life

† Deceased.

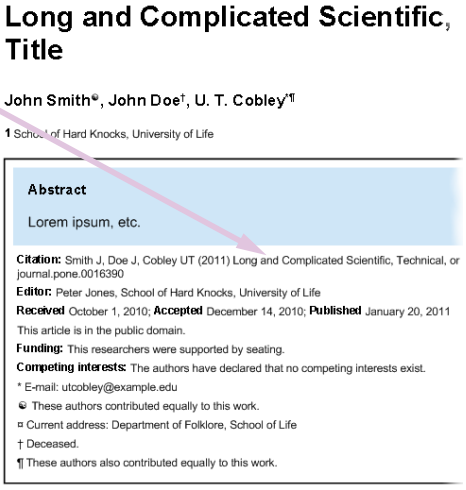
¶ These authors also contributed equally to this work.

Here are some of the elements in the `article-meta` from a sample article alongside the rendering of the article metadata and abstract from the article. I'll be using these to illustrate the sorts of the rearrangements that you need to do to present the information to the reader.

Note that the `abstract` is towards the end of the `article-meta` in JATS, so there's already some selecting and rearranging going on for you to see it in its position on the page.

Reusing `article-title` 65

```
<article-title>Long and Complicated...
...
<contrib deceased="yes">
  <name>
    <surname>Doe</surname>
    <given-names>John</given-names>
  </name>
  <xref ref-type="aff" rid="aff1">
    <sup>1</sup>
  </xref>
</contrib>
<pub-date pub-type="epub">
  <day>20</day><month>1</month><year>2011</year>
</pub-date>
<volume>6</volume><issue>1</issue>
...
<date date-type="received">
  <day>1</day><month>10</month><year>2010</year>
</date>
<date date-type="accepted">
  <day>14</day><month>12</month><year>2010</year>
</date>
```



Long and Complicated Scientific, Title

John Smith☯, John Doe†, U. T. Cobley¤¶

1 School of Hard Knocks, University of Life

Abstract

Lorem ipsum, etc.

Citation: Smith J, Doe J, Cobley UT (2011) Long and Complicated Scientific, Technical, or journal.pone.0016390

Editor: Peter Jones, School of Hard Knocks, University of Life

Received October 1, 2010; **Accepted** December 14, 2010; **Published** January 20, 2011

This article is in the public domain.

Funding: This researchers were supported by seating.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: utcobley@example.edu

☯ These authors contributed equally to this work.

¤ Current address: Department of Folklore, School of Life

† Deceased.

¶ These authors also contributed equally to this work.

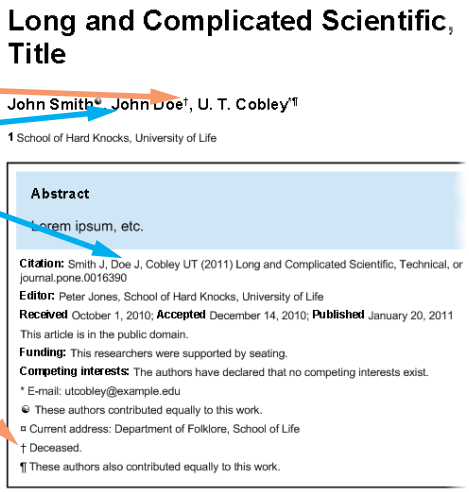
The `article-title` is, of course, presented at the top of the article, but it's also reused when multiple parts of the `article-meta` are pulled together to make the citation.

Either the `article-title` or an `alt-title` is also used for the running head at the top of each of the following pages.

© 2014 Mentea

Reusing and Remodelling Authors


```
<article-title>Long and Complicated...
...
<contrib deceased="yes">
  <name>
    <surname>Doe</surname>
    <given-names>John</given-names>
  </name>
  <xref ref-type="aff" rid="aff1">
    <sup>1</sup>
  </xref>
</contrib>
<pub-date pub-type="epub">
  <day>20</day><month>1</month><year>2011</year>
</pub-date>
<volume>6</volume><issue>1</issue>
...
<date date-type="received">
  <day>1</day><month>10</month><year>2010</year>
</date>
<date date-type="accepted">
  <day>14</day><month>12</month><year>2010</year>
</date>
```



The information about each of the contributors – who could have roles other than just author – is in the XML in a form that's easy to check.

The names are presented one way in the list of authors and another way in the citation, and neither way is exactly the same as in the XML. In the authors' names in the generated citation, the text of each `given-name` has been processed to be just the initials without any punctuation.

The "`deceased="yes"`" for John Doe triggers both the "†" after the name and the inclusion of either a specific "deceased" footnote from the `article-meta`, if it exists, or the standard text note for a deceased author if it doesn't.

You can't see it from the screenshot, but in the PDF, the superscript numerals after the author names are each a link to the corresponding affiliation.

© 2014 Mentea

Reusing and Rearranging Dates

67

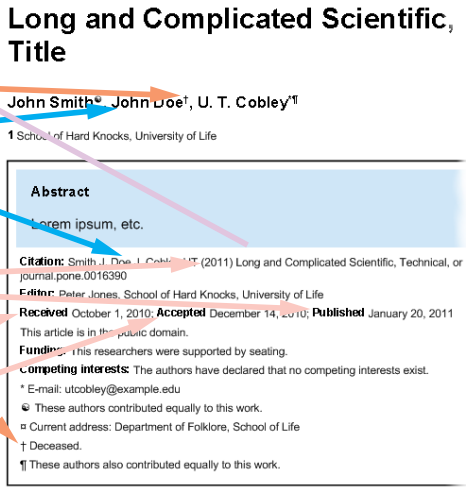
```
<article-title>Long and Complicated...
...
<contrib deceased="yes">
  <name>
    <surname>Doe</surname>
    <given-names>John</given-names>
  </name>
  <xref ref-type="aff" rid="aff1">
    <sup>1</sup>
  </xref>
</contrib>
<pub-date pub-type="epub">
  <day>20</day><month>1</month><year>2011</year>
</pub-date>
<volume>6</volume><issue>1</issue>
...
<date date-type="received">
  <day>1</day><month>10</month><year>2010</year>
</date>
<date date-type="accepted">
  <day>14</day><month>12</month><year>2010</year>
</date>
```

As you might expect by now, the dates for when the article was received, accepted, and published are in the XML in a form that's useful for checking and processing, and are presented in a form that's meant for reading. And only the `year` from the `pub-date` is used in the generated citation.

Putting It All Together

68

```
<article-title>Long and Complicated...
...
<contrib deceased="yes">
  <name>
    <surname>Doe</surname>
    <given-names>John</given-names>
  </name>
  <xref ref-type="aff" rid="aff1">
    <sup>1</sup>
  </xref>
</contrib>
<pub-date pub-type="epub">
  <day>20</day><month>1</month><year>2011</year>
</pub-date>
<volume>6</volume><issue>1</issue>
...
<date date-type="received">
  <day>1</day><month>10</month><year>2010</year>
</date>
<date date-type="accepted">
  <day>14</day><month>12</month><year>2010</year>
</date>
```



I've shown you only some of the reordering and transforming that goes on to present the article metadata, but you can see that it's quite possible to rearrange what's in the XML into the format that you want to present to the reader.

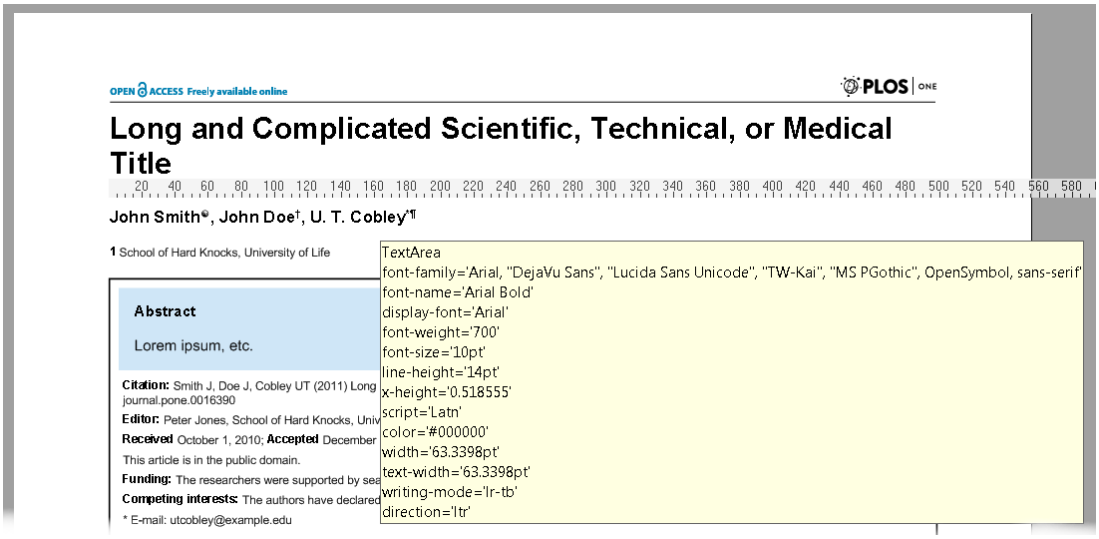
© 2014 Mentea

Checking Output 69

- Work with PDF
  - Ruler in PDF reader or Illustrator
  - Export as RTF (YMMV)
  - Paste into word processor to check fonts and sizes (YMMV)
- Open in FO processor application
  - Antenna House
- xmlroff testing module
- Antenna House Regression Testing Framework (AHRTS)

Checking your output isn't a separate phase, since you should be doing this all the way through.

AHF GUI Preview 70



The Antenna House GUI application can be useful for showing text properties and for measuring distances. It would be even more useful, however, if it could also show the formatted sizes of areas.

© 2014 Mentea

xmlroff Testing Module 71

- Works with any command-line XSL processor
- Compares current results against reference
- Summary and individual HTML reports
- “Stereo” view of differences

PLOS ONE testsuite results

Full	Issues	Total
6	0	6
100%	0%	100%

0 tests with differences

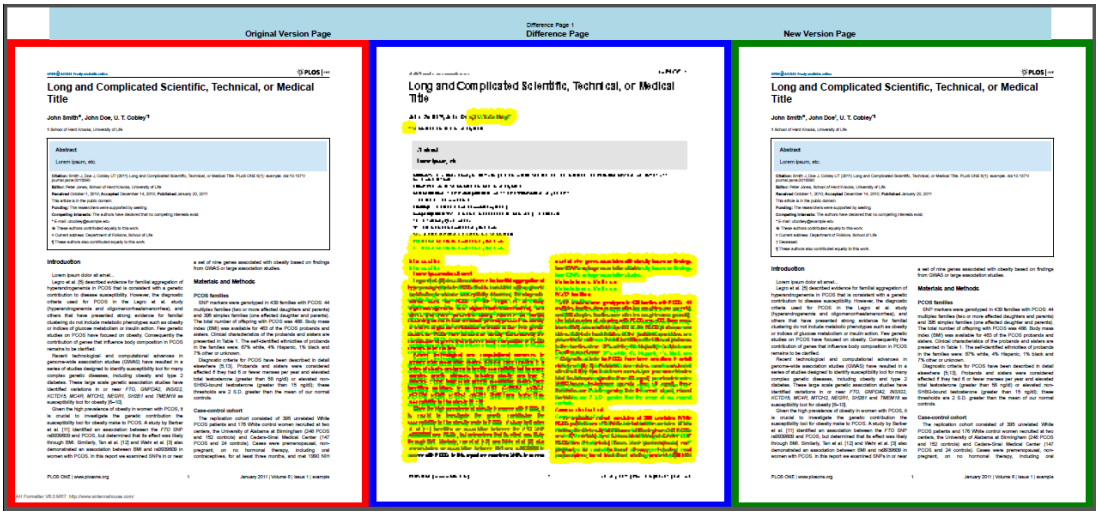
xmlroff Test Suite [Top]

PONE [Top]

pone.0047568	Full	PDF 8
pone.0055261	Full	PDF 11
pone.0055305	Full	PDF 9
pone.0055510	Full	PDF 7
test.table-overflow-height	Full	PDF 17
test.table-page-high	Full	PDF 18

I like to use the xmlroff Testing Module, since I wrote it.

Antenna House Regression Testing System (AHRTS) 72



AHRTS is able to run on two PDFs or two directories of PDFs and produce a PDF report of differences between the files.

© 2014 Mentea

Summary: PLOS ONE 73

- Hard but doable
- Column-wide floats require vendor extension
- Needed multiple passes and access to area tree
- Navigating area tree isn't easy
- No standard for area tree XML made it harder and even less portable
- Creating new FO and reprocessing easier than rewriting area tree
- EXPath Binary Module (and a TIFF-handling library!) could avoid using ImageMagick
  - Or use vendor extension

Questions? 74

質問がありますか。

> それで以上です。どうもありがとうございました。質問がありますか。

References 75

- slide 4 – https://en.wikipedia.org/wiki/Hanabusa_Itch%C5%8D#mediaviewer/File:Blind_monks_examining_an_elephant.jpg
- slide 4 – https://en.wikipedia.org/wiki/Blind_men_and_an_elephant
- slide 5 – CSS ... won on the web http://lists.w3.org/Archives/Public/www-style/2012Sep/0459.html
- slide 5 – Much of it probably from XSL FO http://www.biglist.com/lists/lists.mulberrytech.com/xsl-list/archives/201311/msg00010.html
- slide 18 – http://markmail.org/message/cyvobcwzdyed5ier
- slide 19 – http://lists.w3.org/Archives/Public/public-ppl/2013Dec/0038.html

© 2014 Mentea

Appendix A
About

Tony Graham（トウニー　グラハム）のプロファイル *39*
Mentea（メンテア）会社のプロファイル *39*

Tony Graham（トウニー　グラハム）のプロファイル

トウニーグラハムは 1991 年よりマークアップ言語、1996 年より XML、1998 年より XSLT／XSL-FO に取り組む。 旧 XPPL（XSL-FO 仕様策定のための W3C の XML 印刷・ページレイアウト WG）に参加。現在は W3C の印刷とページレイアウト・コミュニティ議長、著名な XSLT 専門家、`xmlroff` という XSL-FO プロセサを開発、“Unicode: A Primer”の著者です。さらに XSLT の XSpec と Juxy テスチング続けて色々なトレーニングをしています。

トウニーの XML と SGML の職業歴史は日本、アメリカ、イギリス、アイルランドを含む。英語や中国語や日本語や韓国語など学問・自動車・出版・ソフトウェア・通信に利用するということです。 XML、XSLT、XSL-FO、EPUB、色々なテクノロジーについて北アメリカやヨーロッパやオーストラリアな会議でお客さんとも話しあいました。

Tony Graham has been working with markup since 1991, with XML since 1996, and with XSLT/XSL-FO since 1998. He is Chair of the Print and Page Layout Community Group at the W3C and previously an invited expert on the W3C XML Print and Page Layout Working Group (XPPL) defining the XSL-FO specification, as well as an acknowledged expert in XSLT, developer of the open source xmlroff XSL formatter, a committer to both the XSpec and Juxy XSLT testing frameworks, the author of “Unicode: A Primer”, a member of the XML Guild, and a qualified trainer.

Tony’s career in XML and SGML spans Japan, USA, UK, and Ireland, working with data in English, Chinese, Japanese, and Korean, and with academic, automotive, publishing, software, and telecommunications applications. He has also spoken about XML, XSLT, XSL-FO, EPUB, and related technologies to clients and conferences in North America, Europe, and Australia.

Mentea（メンテア）会社のプロファイル

MENTEA は XML や XSL-FO や XSLT など相談とトレーニングを提供しています。国際的の会議とトレーニングに加えて、電子メールやスカイプやインスタント・メセージや電話などでお客さんと連絡を保つ。安全な wiki トラキングシステムも利用してしています。

MENTEA のスタッフは 1991 年よりマークアップ言語、1996 年より XML、1998 年より XSLT／XSL-FO に取り組む。アイルランドのダブリンに位置しているのですが、世界的なお客さんを支えています。アメリカ・アイルランド・イギリス・フランスの会社と団体を助力て、色々なサービスを提供する。

MENTEA のサービス：

- XSLT, XSL, XML トレーニング。
- プロフェッショナルの団体のために Schematron を書くという。
- XSLT で作った自動的な schema 書類システム（HTML と PDF を制作したシステム）を増えるという。
- ソフトウェア会社のために FOP を発展する。
- XML、oXygen、DocBook、XSLT 2.0、XSL-FO のトレーニング。
- 科学のジャーナルのために JATS のフォーマットは PDF に変わるという。

© 2014 Mentea

- XSLT スタイルシートを製作して、XML ではないスタイルシートは XML に変えて、また EPUB に変える。
- 商業銀行のために XSLT を製作して EXCEL は XML にかえります。

MENTEA は特別の特殊な技術を供給する。XML と XSL-FO/XSLT ではなく（Unicode・SGML・ DSSSL・ C プログラム・Java・Perl・ Lisp）色々な言語もできる。

マークアップはよく理解してしています。日本やアメリカやイギリスやアイルランドなどマークアップ利用しています。ユーザー、相談者、開発者として、データは英語・フランス語・中国語・日本語・韓国語で色々な（学問・自動車・出版・ソフトウエー・電話・インタネットサービス・ドキュメントプロセシング）使い方があります。

また、多くの道具（ユニットテステイング・コードわたる・プロファイリング・を使って、ソフトウエーの良質を安全にできます。

複雑か特別なプロジェクトがあるばいは、世界中の相談者達と団体を手伝いを頼む。

Mentea specialises in consulting and training in XML, XSL-FO, & XSLT. We are available for on-site meetings and classes, worldwide, but as well as on-site meetings and classes, we routinely keep in touch with clients though email, Skype, instant messaging, and telephone and through a secure, per-client or per-project wiki, revision-control, and issue-tracking system.

Our staff have been working with markup since 1991, with XML since 1996, and with XSLT/XSL-FO since 1998. Based in Dublin, Ireland, Mentea has a global reach: in recent projects, we have helped companies and organisations in the USA, Ireland, England, and France with their XSLT, XSL, and XML, including:

- Writing Schematron for a professional body
- Augmenting a XSLT-based automated schema documentation system that produces both HTML and PDF
- Extending FOP for a software company
- Training in XML, oXygen, DocBook, XSLT 2.0, and XSL-FO
- Formatting JATS to PDF for a scientific journal
- Writing XSLT stylesheets to convert non-XML into XML then into EPUB
- Writing XSLT to convert Excel into XML for a commercial bank

Mentea presents a unique range of skills extending beyond XML and XSL-FO/XSLT into Unicode, SGML, DSSSL, and programming in C, Java, Perl, Lisp, and other languages.

We understand how markup works. Our staff has worked with markup in Japan, USA, UK, and Ireland as user, consultant, and developer, with data in English, French, Chinese, Japanese, and Korean, with academic, automotive, publishing, software, and telecommunications applications, and in the Web Services and document processing arenas.

We are also interested in applying the tools for ensuring software quality – unit testing, code coverage, profiling, and other tools – to XML and XSLT/XSL-FO processing.

Through our associations and affiliations with other consultants around the world, we can call on extra help for large or specialised projects.

© 2014 Mentea

© 2014 Mentea

MENTEA