mPach
Integrated Publishing and Archiving of Journals in HathiTrust

Seth Johnson, Bryan Smith, & Kevin S. Hawkins
Michigan Publishing
Overview

1. Overview of mPach, a package of tools for publication of born-digital journals in HathiTrust

2. Introduction to mPach’s Prepper interface

3. Technical discussion of mPach’s Norm utility for converting Word DOCX files to JATS XML
Michigan Publishing

Michigan Publishing is the primary academic publisher of University of Michigan and is based in the University Library.

Michigan Publishing has long used a system called DLXS as its primary platform for online content, but we need an architecture that will scale better in order for us to continue to grow.

www.publishing.umich.edu
What is HathiTrust?

Partnership of research libraries around the world

Shared digital repository certified to be preservation-quality with over 11 million digitized volumes (nearly 500 terabytes of data)

www.hathitrust.org
Publishers and Archives

Publishers require flexibility to innovate. But archives need stability.

HathiTrust provides us with an infrastructure in which to provide long-term preservation and discoverability while allowing for innovative services to be built on top.
Main design principle

Archiving happens as a byproduct of publication rather than after the fact.
JATS and mPach

JATS was selected because of the increasing coalescence of the publishing industry around this open, non-proprietary standard.

Publishing ("blue") tag set works for born-digital literature, with a constrained set of tags to render, unlike "green". But unlike "orange" it also includes important metadata elements (in <front>).
mPach overview

www.lib.umich.edu/mpach
Prepper

Dashboard for administering a journal and putting manuscripts through the production process

Guides the conversion process from DOCX to JATS (using Norm)

Ruby on Rails application
## Journal of Electronic Publishing

### Upload article

- **Choose File**: No file chosen
- **Submit**

### Resume article

- **Search**

<table>
<thead>
<tr>
<th>Article Title</th>
<th>Author</th>
<th>Vol</th>
<th>Issue</th>
<th>Status</th>
</tr>
</thead>
</table>

### Manage Volumes

- **Volume**: Foo
- **Publication Year**
- **Volume No.**

<table>
<thead>
<tr>
<th>Issue No.</th>
<th>Publication Date</th>
<th>Issue Title</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notifications**

- **Hypernews, Hyperreaders and Beyond** was accepted by HathiTrust yesterday
- **Hypernews, Hyperreaders and Beyond** was submitted to HathiTrust three days ago
- **Kevin Hawkins** uploaded article *Two Future Binaries* to *Journal of Electronic Publishing* six days ago
- **Kevin Hawkins** can now submit articles to *Journal of Electronic Publishing* two weeks ago
**Prepper Article Prep (2 of 8)**

**JOURNAL OF ELECTRONIC PUBLISHING**

**COLOR VARIABILITY AND BODY SIZE OF LARVAE OF TWO EPOMISSPECIES (COLEOPTERA, CARABIDAE) IN ISRAEL, WITH A KEY TO THE LARVAL STAGES**

---

**Conversion Report**

**Article Metadata**

Please review the following metadata. If there are any typos or excluded fields, please correct in your source document and re-submit.

- **ARTICLE TITLE**: Color variability and body size of larvae of two Epomisspecies (Coleoptera, Carabidae)
- **ARTICLE SUBTITLE**:
- **AUTHORS**:
  - Given name: Gil
  - Surname: Wizen
  - Given name: Avital
  - Surname: Gashith
- **CONTRIBUTORS**:
  - No contributors found
- **KEYWORDS**: Epomis larvae, Carabidae, color atlas, body size
- **ABSTRACT**: Species identification using the characteristics of developmental stages is challenging. However, for insect taxonomy the coloration of larval stages can be an informative feature. The use of live specimens is recommended for this because the color fades in preserved specimens. In this study we examine the
- **PUB DATE**:

---

**mPach**

**Dashboard**

**Journals**

**Help**

**feedback**
### Add Media

The following 9 media files were detected. Please review.

<table>
<thead>
<tr>
<th>Media</th>
<th>Filename</th>
<th>Size</th>
<th>Alt text</th>
<th>Upload new version</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image6.jpg" alt="Image" /></td>
<td>image6.jpg</td>
<td>340 KB</td>
<td></td>
<td>Choose File  No file chosen</td>
</tr>
<tr>
<td><img src="image9.jpg" alt="Image" /></td>
<td>image9.jpg</td>
<td>95 KB</td>
<td></td>
<td>Choose File  No file chosen</td>
</tr>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td>image1.jpg</td>
<td>173 KB</td>
<td></td>
<td>Choose File  No file chosen</td>
</tr>
</tbody>
</table>
Color variability and body size of larvae of two *Epomis* species (Coleoptera, Carabidae) in Israel, with a key to the larval stages

Gil Wizen
Avital Gaash

Abstract
Species identification using the characteristics of developmental stages is challenging. However, for insect taxonomy the coloration of larval stages can be an informative feature. The use of live specimens is recommended for this because the color fades in preserved specimens. In this study we examine the possibility of using variation in coloration and color pattern of larvae in order to distinguish between two ground beetles species *Epomis degeeri* (Dejean, 1831) and *Epomis circumscriptus* (Duftschmid, 1812). We present an atlas and describe the coloration and body size of the three larval stages of the above species based on live specimens. An identification key is given for the three larval instars of the two *Epomis* species. The first instar larva of the two *Epomis* species can be easily distinguished based on their color. From the second instar on, the variability in coloration and color patterns increases, creating an overlap in these attributes between larvae of the two species. Except for minor differences in color of the antennae and the base of the mandibles, larvae of the two species are indistinguishable at the second
Review

We successfully build and validated the METS package. This article is ready for submission to HathiTrust.
Confirmation of submission

mPach submission # 0123456789

The HathiTrust has received your submission for

Color variability and body size of larvae of two Epomis species (Coleoptera, Carabidae) in Israel, with a key to the larval stages

Return to journal...
Refurbishing the Camelot of Scholarship: How to Improve the Digital Contribution of the PDF Research Article

John Willinsky, Alex Garrett, and Angela Pan Wong

This paper was refereed by the Journal of Electronic Publishing’s peer reviewers.

Abstract

The Portable Document Format (PDF) has become the standard and preferred form for the digital edition of scholarly journal articles. Originally created as a solution to the need to “view and print anywhere,” this technology has steadily evolved since the 1990s. However, its current use among scholarly publishers has been largely restricted to making research articles print-ready, and this greatly limits the potential capacity of the PDF research article to form a greater part of a digital knowledge ecology. While this article considers historical issues of design and format in scholarly publishing, it also takes a very practical approach, providing demonstrations and examples to assist publishers and scholars in finding greater scholarly value in the way the PDF is used for journal articles. This involves but is not limited to graphic design and bibliographic linking, the deployment of metadata and research data, and the ability to combine elements of improved machine and human readability.

Introduction

The Portable Document Format (PDF) was released by Adobe Systems in 1993 to facilitate the electronic distribution of documents. It was created to assist the circulation of digital documents among the newly networked computers that were spreading throughout offices, whether in local area networks (LAN) or through the internet. What had become apparent was that documents were being prepared by various word-processing programs, each with their own proprietary file format. With networking Rate racing ahead of file compatibility, John Warnock, Adobe Systems co-founder, in 1991 initiated what he called the Camelot Project in order to solve the “view and print anywhere” problem, as he neatly characterized it (1995, p. 1). Nearly a decade earlier, in 1982, the resourceful Warnock, working with Charles Geschke, figured they had solved the same problem with PostScript (marking the beginning of Adobe Systems).

However, PostScript was itself not proving universally applicable. It required “powerful desktop machines,” as Warnock put it, as well as PostScript printers (1991, pp. 1–2). The goal of Camelot was to develop a lightweight file format that would serve the broadest possible range of users, at least until widespread computing power caught up with the demands of PostScript. Camelot was intended, then, as a temporary, transitional solution to the view-and-print anywhere problem. Its history and success proved otherwise. When launched in 1993, the file format’s poetic Camelot moniker was replaced by the prosaic “portable document format,” now universally known as PDF. In 2008, Adobe released the PDF as an open standard for others to develop applications for writing and reading it, in what we might think of as the new twenty-first-century corporate spirit of open standards and open source software.

In scholarly communication, the PDF has become the standard file format for research articles published in the electronic edition of peer-reviewed journals. Although many journals also publish a HTML version of their articles along with a PDF, the bulk of the research literature is now available in PDF. Over the last decade, the majority of researchers have switched to reading the online edition of journals available through their library’s electronic collections. (King, Tenopire, Cheomprayong, and Wu, 2009, p. 173; Hemminger, Lu, Vaughn, and Adams, 2007). While finding articles online is becoming a common practice, most academic faculty print out a good proportion of the PDFs they wish to read, while


In scholarly communication, the PDF has become the standard file format for research articles published in the electronic edition of peer-reviewed journals. Although many journals also publish a HTML version of their articles along with a PDF, the bulk of the research literature is now available in PDF. Over the last decade, the majority of researchers have switched to reading the online edition of journals available through their library’s electronic collections. (King, Tenopire, Cheomprayong, and Wu, 2009, p. 173; Hemminger, Lu, Vaughn, and Adams, 2007). While finding articles online is becoming a common practice, most academic faculty print out a good proportion of the PDFs they wish to read, while
Journal View in HathiTrust
Norm

Converts DOCX to JATS XML:

1. Parse DOCX XML

2. Internal Representation and Mapping

3. Create JATS XML and assets
Norm Usage

Stand-alone command-line application

Input: DOCX or ODT file

Output:

document_name.zip/
document_name.xml (JATS)
assets/
    image_1.png
    image_2.png
Color variability and body size of larvae of two *Epomis* species (Coleoptera, Carabidae) in Israel, with a key to the larval stages

Species identification using the characteristics of developmental stages is challenging. However, for insect taxonomy the coloration of larval stages can be an informative feature. The use of live specimens is recommended for this because the color fades in preserved specimens. In this study we examine the possibility of using variation in coloration and color pattern of larvae in order to distinguish between two ground beetles species *Epomis decemnotatus* (Reichel, 1831) and *Epomis circumscriptus* (Druce, 1812). We present an atlas and describe the coloration and body size of the three larval stages of the above species based on live specimens. An identification key is given for the three larval instars of the two *Epomis* species. The first instar larvae of the two *Epomis* species can be easily distinguished based on their color. From the second instar on, the variability in coloration and color patterns increases, creating an overlap in these attributes between larvae of the two species. Except for minor differences in the color of the antennae and the base of the mandibles, larvae of the two species are indistinguishable at the second and third larval stages. To the best of our knowledge this is the first attempt to use variation in coloration and color pattern in live larvae in order to identify coleopterans. The color atlas of the larvae enables simple separation of the two *Epomis* species without requiring sophisticated magnifying devices, although it is less straightforward at the second and third larval stages. We found similar body lengths between the two species for all developmental stages, except for third instar larvae prior to pupation. In the two species the difference in larval body length before pupation positively correlated with that of the adult beetles. More than 70% of the adults length can be explained by the length of the late third-instar larva; i.e. the large larvae develop into large adults. The larger specimens are the females.

*Epomis* larva, Carabidae, color atlas, body size

**Introduction**

Coloration can be an informative feature in insect taxonomy (van Emden 1957, Luff 1993). While the larvae of some ground beetles (Carabidae) have been well studied (reviewed in Lawrence 1991) for other beetles the larval stage is still unknown. Those
Norm Transformation Process

Given:

Word document

Configuration specifying:
- Word styles corresponding to each JATS element
- Parents for each JATS element
- Appropriate section (head, body, back) for each JATS element
Step 1: Transform data into internal representation

Create empty array for each section (front, body, back)

For each element in DOCX body:
- Find style and contents of element
- Determine which JATS element (configuration)
- Determine which section (configuration)
- Append tuple [JATS element, content, style] to section's array
Color variability and body size of larvae of two Epomis
Norm configuration mappings (default.cfg)

[ FRON T ]
ArticleTitle = article-title

[ FRONT-PARENTS ]
article-title = title-group
title-group = article-meta
article-meta = front
Title: Color variability and body size of larvae of two *Epomis* species (Coleoptera, Carabidae) in Israel, with a key to the larval stages

In Norm’s internal representation:

('article-title',
 [('Color vari...of two', None, None),
  ('Epomis', ['i'], None),
  ('(Coleoptera...stages', None, None)
  ],
 'ArticleTitle')
Step 2: Render JATS output from internal representation

Create empty Document Object Model (DOM) tree

For each section (front, body, back):
- Add node for section to tree
- For each tuple for section (see step 1):
  - Create node for JATS element tuple
  - Find parent for element (configuration)
  - Attach node to parent

Marshall output to XML.
An article title in JATS

<article>
  <front>
    <title>
      <article-meta>
        <title-group>
          <article-title>
            Color variability and body size of larvae of two <i>Epomis</i> species (Coleoptera, Carabidae) in Israel, with a key to the larval stages
          </article-title>
        </title-group>
      </article-meta>
    </title>
  </front>
</article>
Future Plans for Norm

The <body> of the article is where we’re seeing the most feature creep, making configuration and styles increasingly complicated.

Options:
1. “Norm lite” for the front, another tool (meTypeset) for the body
2. Norm to handle both front and body, refactor of the codebase needed
www.lib.umich.edu/mpach