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 publishing platform, but we need an architecture that will scale better in order for us to continue to grow.

publishing.umich.edu

What is HathiTrust?

Partnership of research libraries around the world

Shared digital repository certified to be preservation-quality with over 10 million digitized volumes (nearly 500 terabytes of data)

HATHI TRUST

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

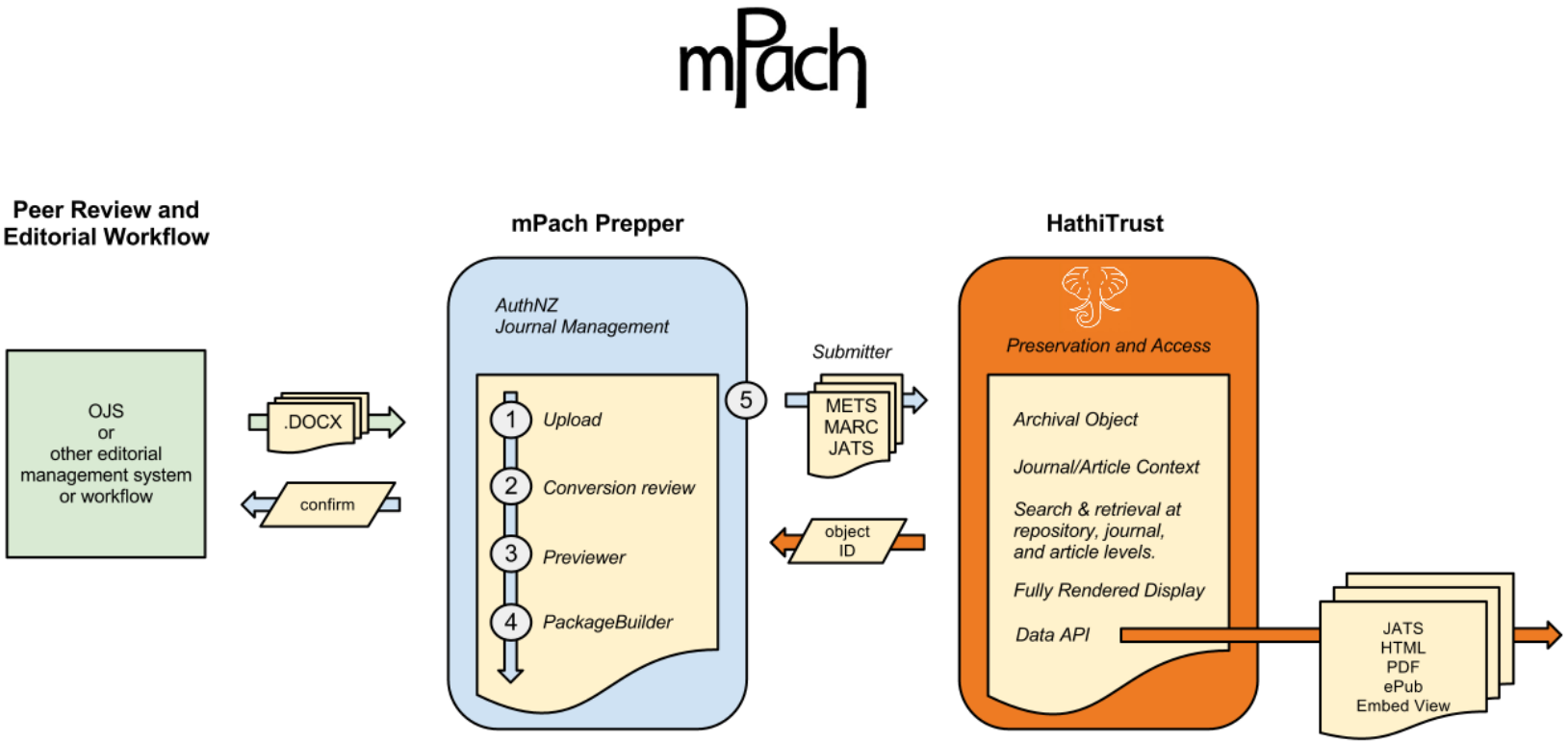
Archiving happens as a byproduct of publication rather than after the fact.

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.hathitrust.org/mpach

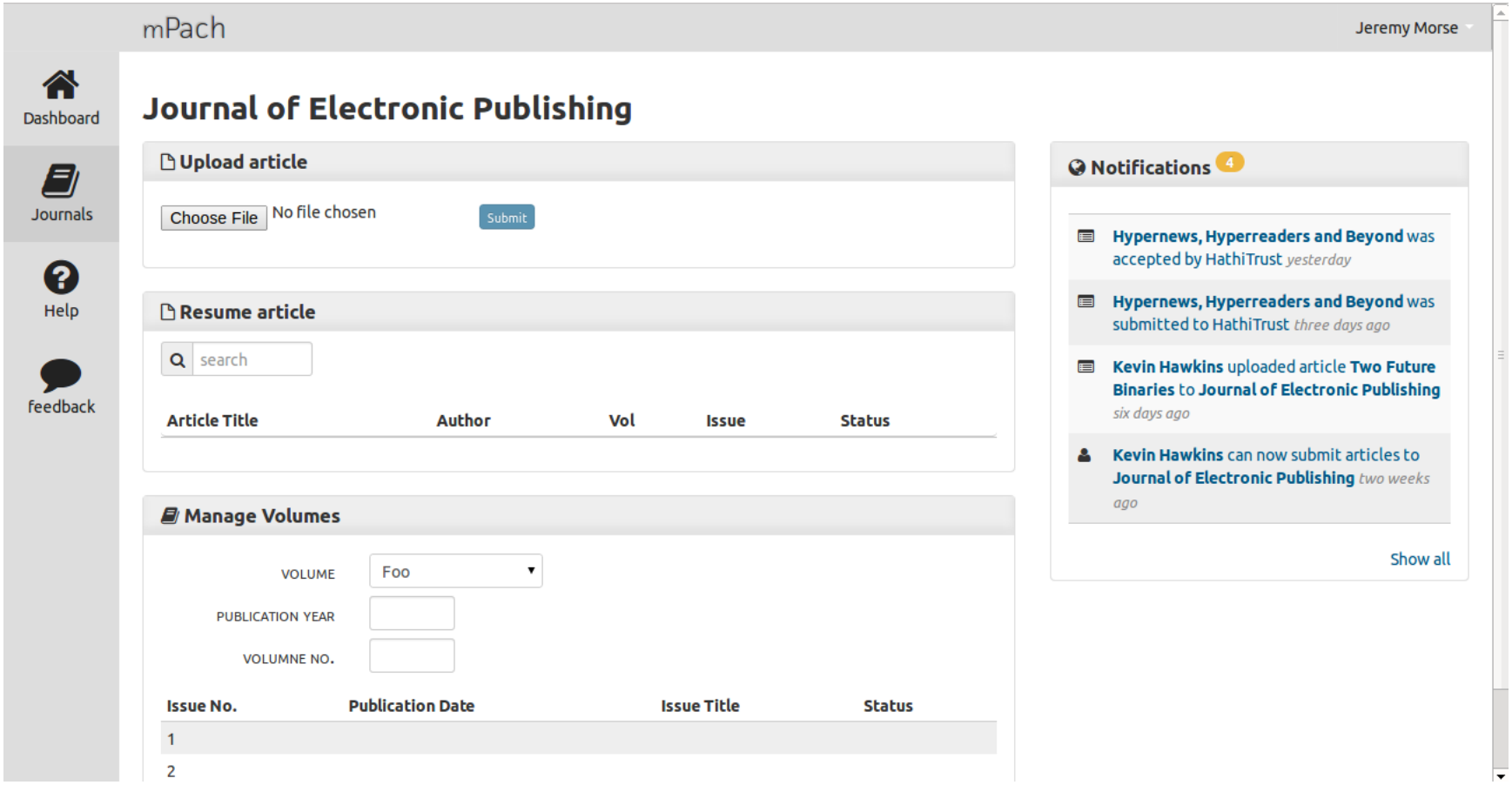
Prepper

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

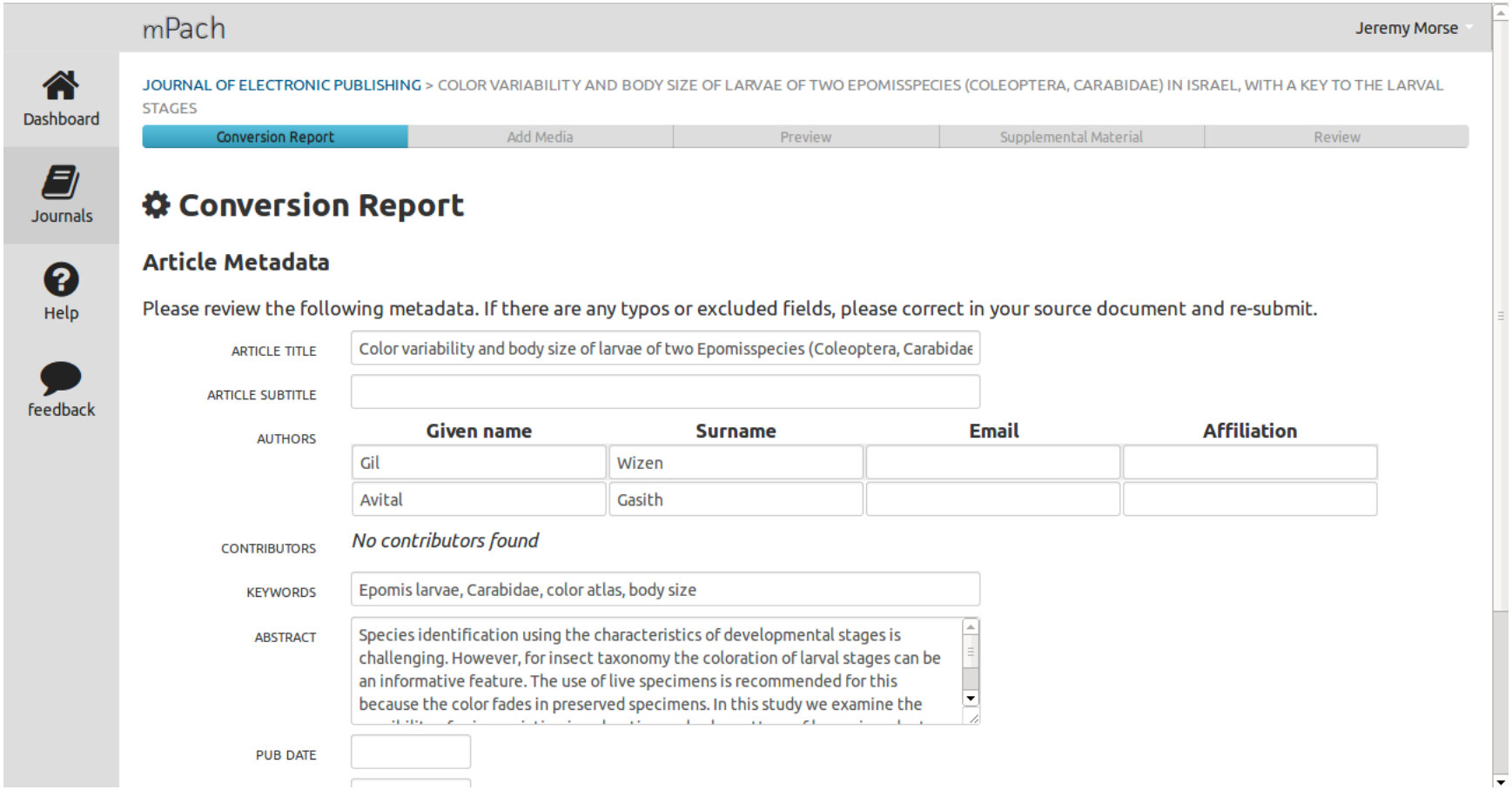
Guides the conversion process from DOCX to JATS (using Norm)

Ruby on Rails application

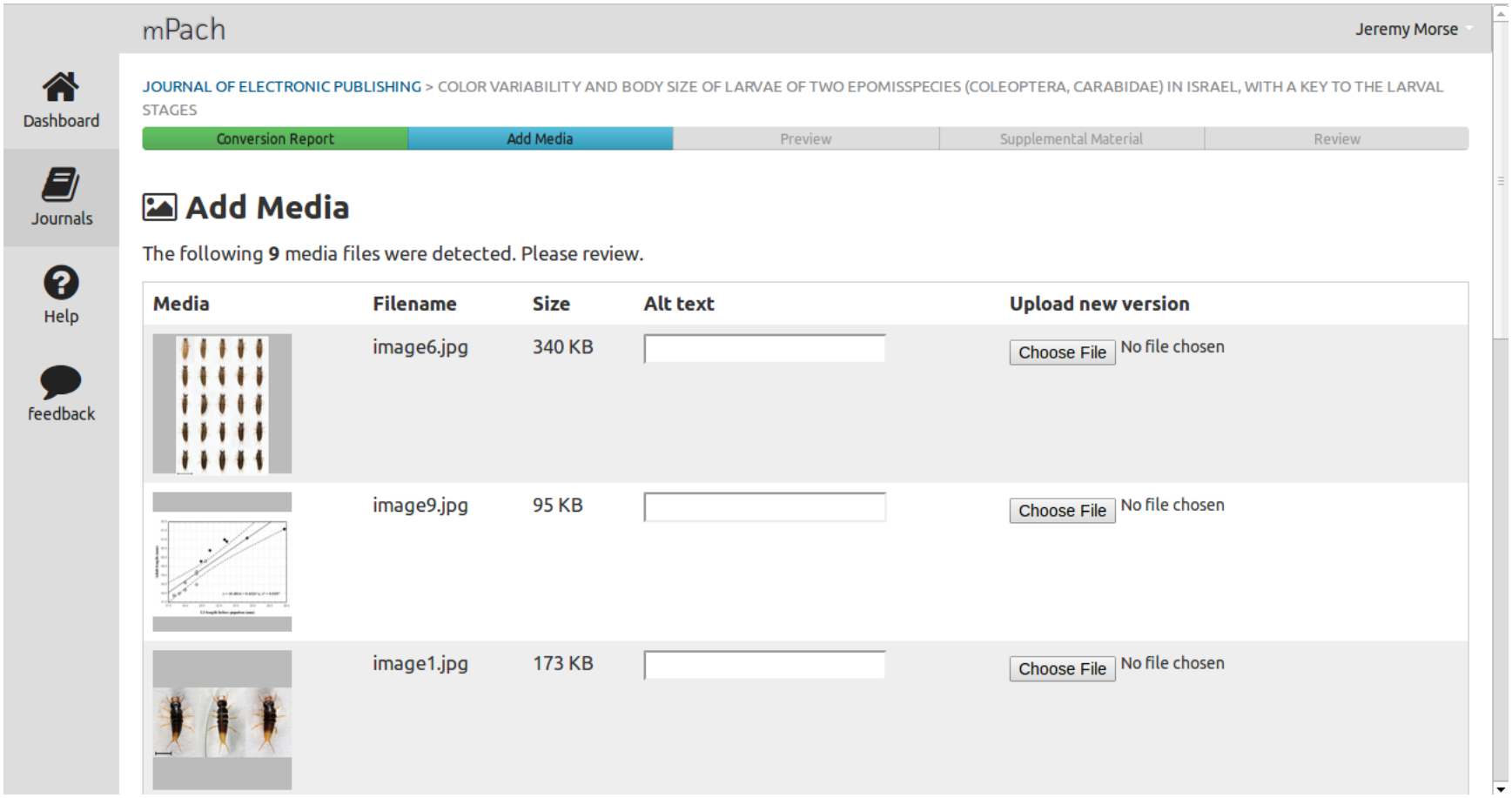
Prepper Article Prep (1 of 8)



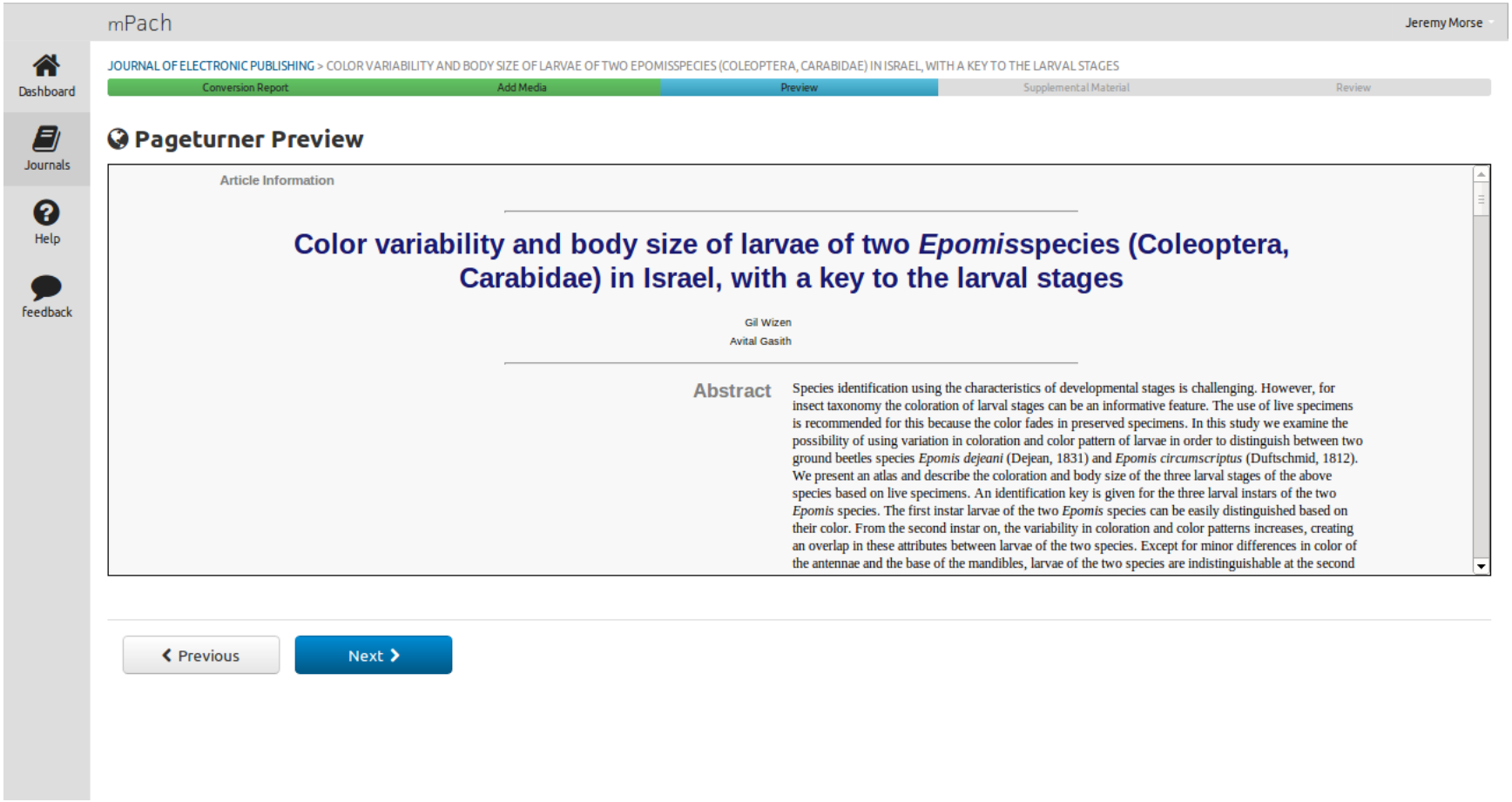
Prepper Article Prep (2 of 8)



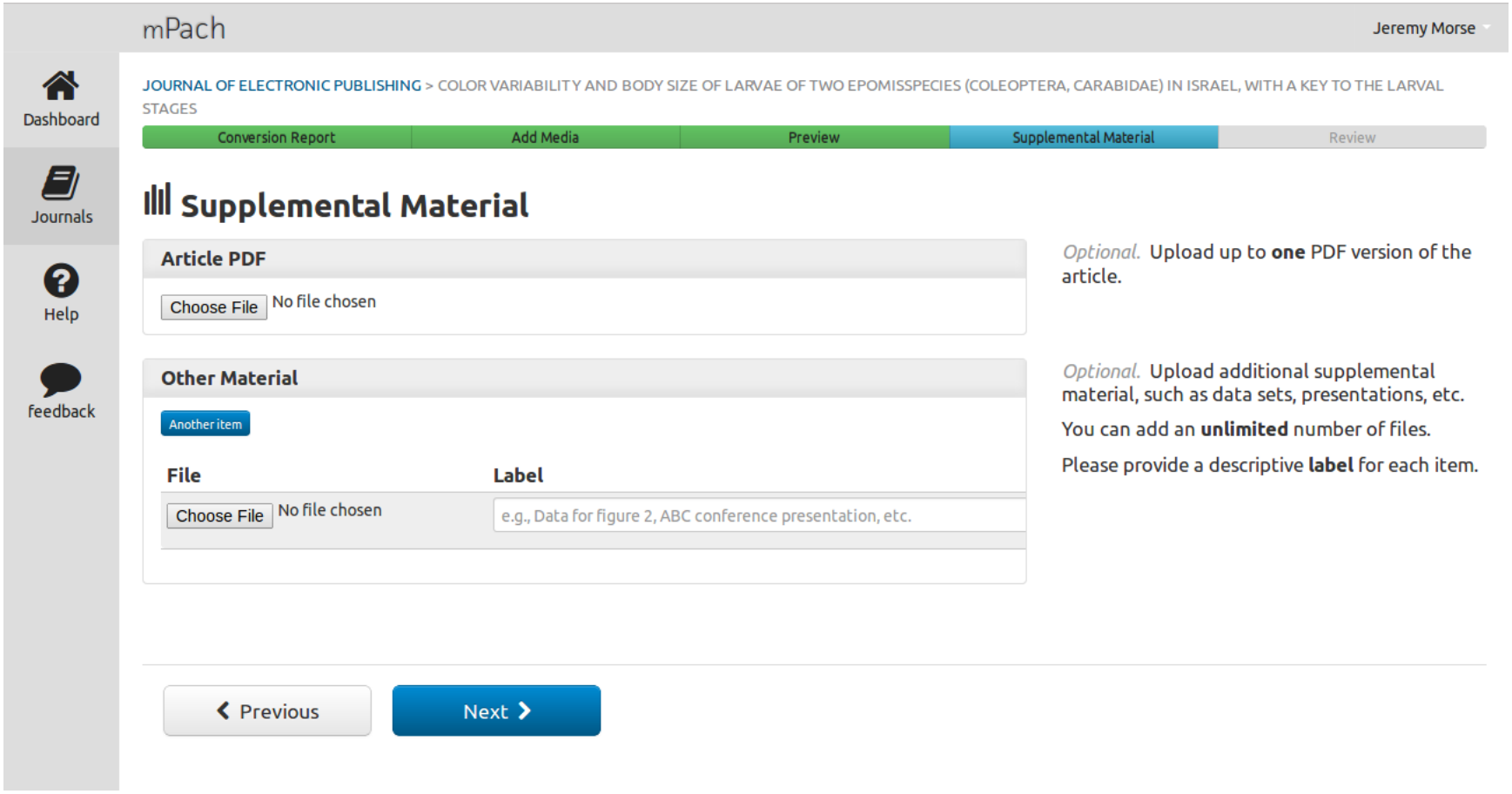
Prepper Article Prep (3 of 8)



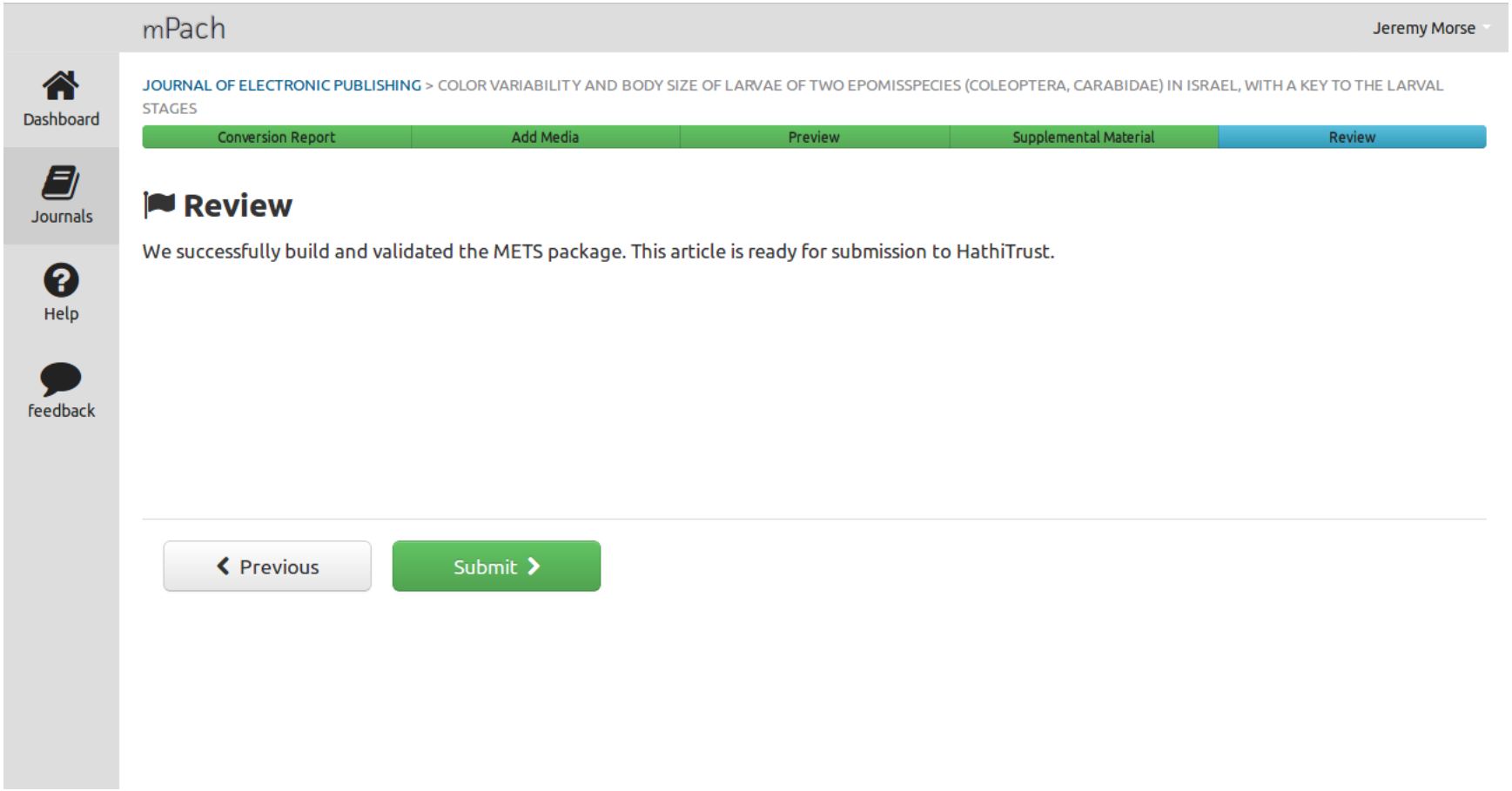
Prepper Article Prep (4 of 8)



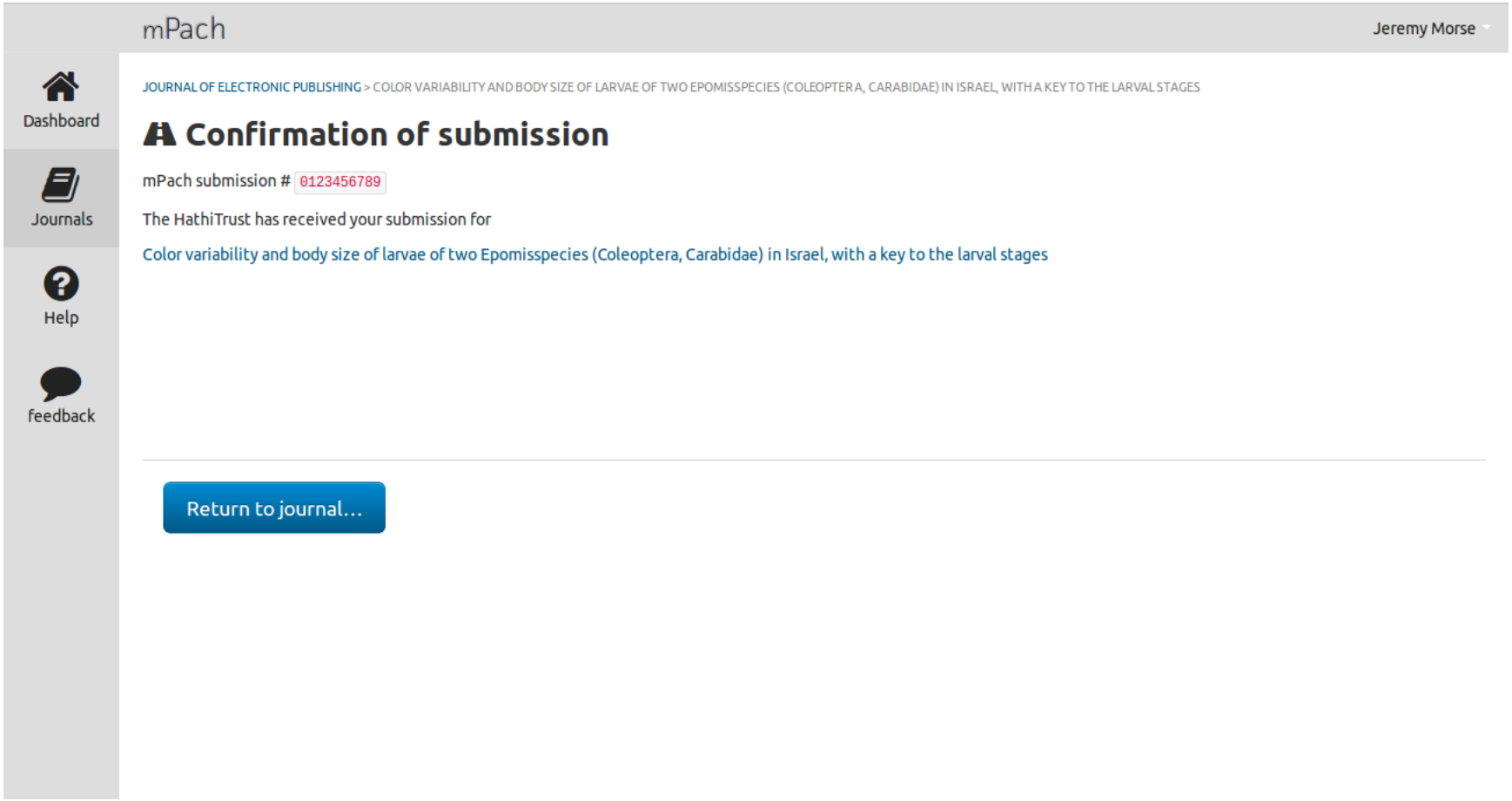
Prepper Article Prep (5 of 8)



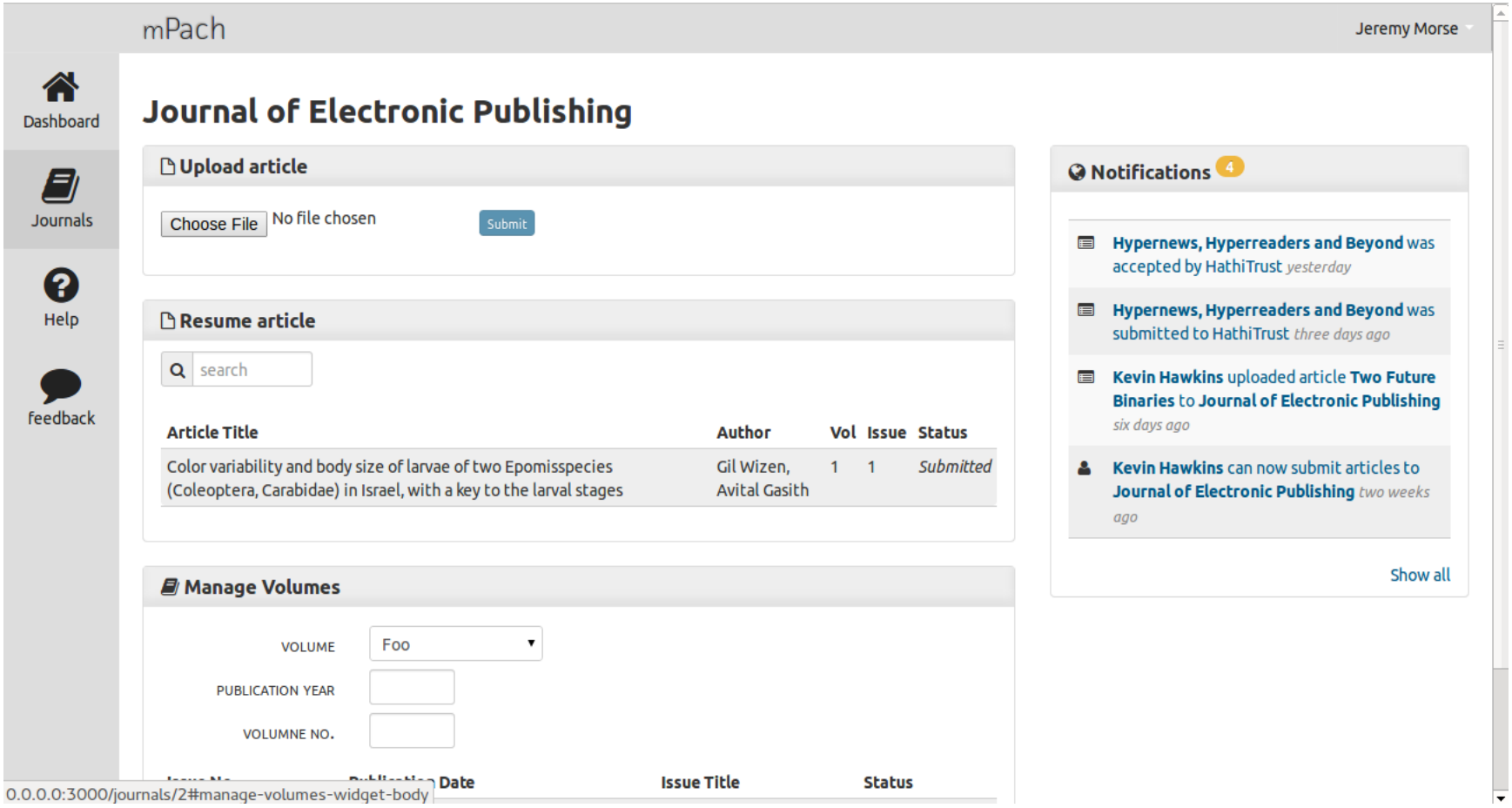
Prepper Article Prep (6 of 8)



Prepper Article Prep (7 of 8)

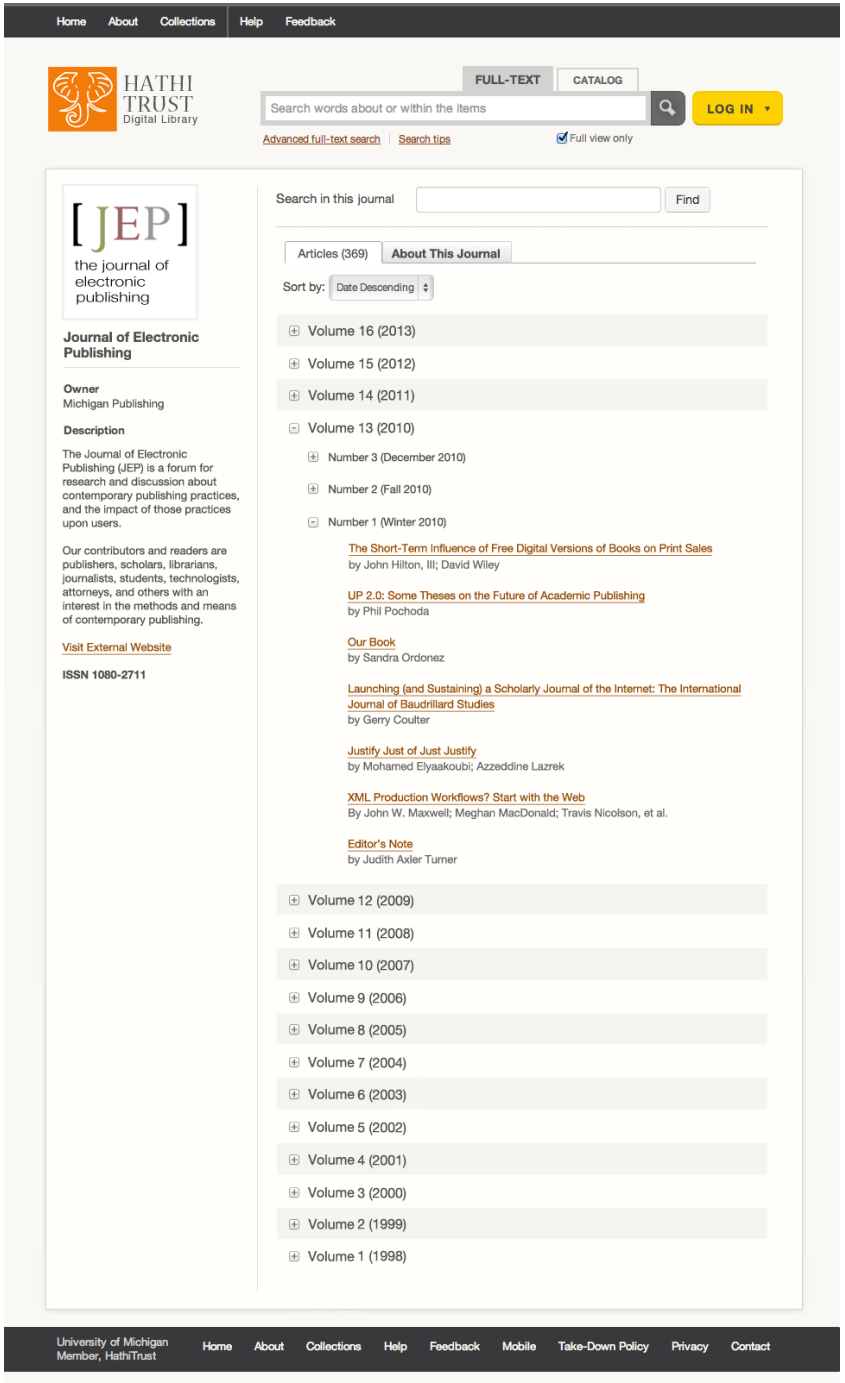


Prepper Article Prep (8 of 8)



Article View in HathiTrust

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

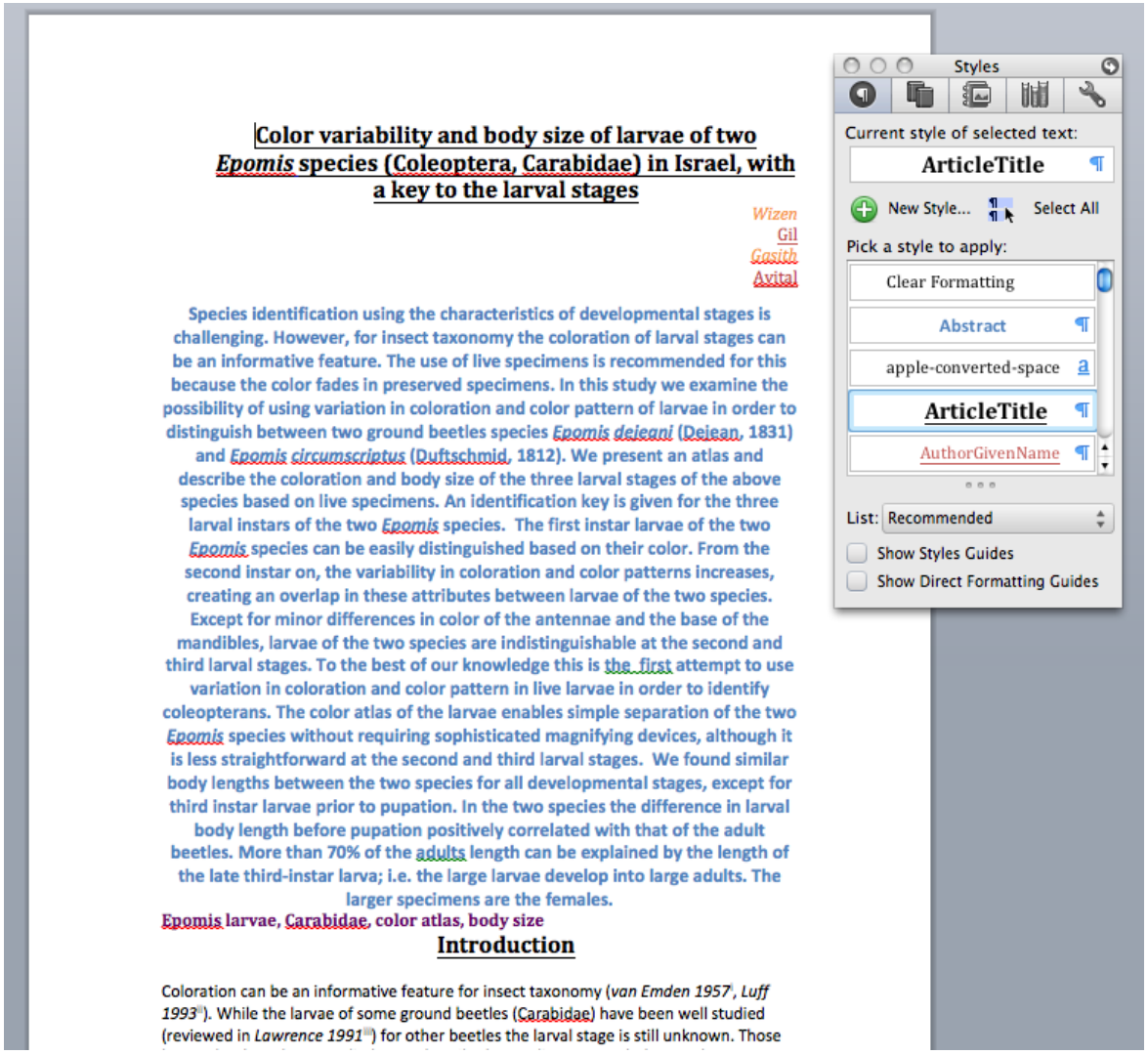
Stand-alone command-line application.

Input: DOCX or ODT file

Output:

- document_name.zip/
  - document_name.xml (JATS)
  - assets/
    - image_1.png
    - image_2.png

Word Styles and Norm



Norm Transformation Process

Given:

Word document

Configuration specifying:

- Word styles corresponding to each JATS elements
- Parents for each JATS elements
- Appropriate section (head, body, back) for each JATS element

Step 1: Transform data into internal representation

Create empty array for each section (head, 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

DOCX XML with Word Style

```
<w:body>
  <w:p>
    <w:pPr><w:pStyle w:val="ArticleTitle"/></w:pPr>
      <w:r>
        <w:t>Color variability and body size of larvae of
two</w:t>
     </w:r>
     <w:r>
       <w:rPr><w:i/></w:rPr>
       <w:t>Epomis</w:t>
    </w:r>
    <w:r>
    ...
```

Norm configuration mappings (default.cfg)

```
[ FRONT ]
ArticleTitle = article-title

[ FRONT-PARENTS ]
article-title=title-group
title-group = article-meta
article-meta = front
```

Step 2: Render JATS output from internal representation

Create empty Document Object Model (DOM) tree

For each section (head, 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.

Sample internal representation: article title

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')
```

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>
```

Future Plans for Norm

Support for OpenDocument (ODF) format underway

Considering other formats (LaTex, PDF)

Planning to release Norm under open source license

www.lib.umich.edu/mpach