Scripture Burrito 0.1 RFC

Mark Howe

May 28, 2019

CONTENTS:

CHAPTER

ONE

THE ROOT ELEMENT

1.1 PROPOSED CHANGES FOR 0.1

- Root element name becomes **BurritoMetadata**, in the null namespace.
- @version must be 0.1
- @id regex should be expanded to include a prefix and to allow other formats of id.
- @revision should be expanded to allow Mercurial and Git commit ids as well as DBL-style positive integers.
- Zero or one id/revision may be placed in the root element, with the others in systemId

1.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

TWO

IDS

2.1 PROPOSED CHANGES FOR 0.1

2.1.1 Required idServer declarations

These appear as the first children of the root element, eg

```
<idServer prefix="dbl">https://thedigitalbiblelibrary.org</idServer>
<idServer default="true">http://atlantisbibleconsortium.net</idServer>
<idServer prefix="myServer" local="true">http://localhost::8080</idServer>
```

- At least one idServer element is required
- The element must contain either a 'prefix' or a 'default="true"' attribute. (It may contain both, and only one idServer may be the default.) The default, if present, will be assumed to apply to namespaces with no prefix.
- The element may contain a 'local' attribute. When true, this signifies that the ids are for internal use only, and that they should be stripped before export.
- The enclosed text is a URI. In schema this means "pretty much any string", but using URLs that resolve to a server endpoint would help with discoverability.

If there is no default idServer, all ids in the document must be prefixed.

2.1.2 ID Syntax

(<prefix>::)?<id>

where

- "prefix" is a NCName (an XML name with no colon)
- "id" matches

```
[0-9A-Za-z]([0-9A-Za-z_-]{0,30}[0-9A-Za-z])?
```

ie a string starting and ending with an alphanumeric character and containing alphanumeric characters, hyphens and underscores.

IDs in this format can be tested for prefixedness (?!) by searching for "::", a substring which seems unlikely to occur in any existing id schemes.

2.1.3 Revision Syntax

The non-prefixed ID regex above ought to allow DBL (numeric) and PT/uW (UUID) revision/commit identifiers.

2.1.4 Expose User ID

This should happen anywhere that the metadata refers to a person by name. It probably needs to be optional since it may not be possible to recover this information retrospectively.

2.1.5 Expose License ids

This would be part of the new license subsection.

2.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

THREE

IDENTIFICATION

3.1 TOP-LEVEL FIELDS

3.1.1 PROPOSED CHANGES FOR 0.1

3.1.1.1 Replace english/Local fields with elements qualified by language

eg

```
<name lang="en">...</name>
<name lang="fr">...</name>
```

Affected fields include

- name/nameLocal
- description/descriptionLocal
- abbreviation/abbreviationLocal

3.1.1.2 basedOn

This identifies the snapshot on which the entry is based, which might be from a different ecosystem. This information is potentially useful for forensics. It also provides a mechanism for 3-way diffing of documents when the two deltas are from different ecosystems.

```
<basedOn type="dbl">
   <id>482ddd53705278cc</id>
   <revision>1</revision>
</basedOn>
```

3.1.1.3 Replace scope with canonicalContent

Scope does not have enough options to describe all projects. In addition, it is unclear whether the scope describes the books actually present (impossible with an enum for incremental publishing) or the intended final scope of the project (which is a somewhat existential concept). Including a global canonicalContent section, as currently exists in publications, for a whole entry, provides scope information in a more flexible and transparent way.

3.1.2 ISSUES TO CONSIDER FOR 0.2

None

3.2 systemId

3.2.1 PROPOSED CHANGES FOR 0.1

3.2.1.1 Add DBL type (“dbl”)

This is required to make the document structure orthogonal.

3.2.1.2 Add other known organizations

- Paratext ecosystem (“ptx”)
- Unfolding Word (“uword”)
- Vachan Online (“vachan”)

3.2.1.3 Support x-* types

The systemId type mechanism was created when DBL needed to work with a small number of large ecosystems. Future ecosystems may be small – maybe a national denomination or even one church. It may not always make sense to add such organizations to the schema and, when it does, this will take some time. Some architectures involve local servers (on a VPN, an intranet or even localhost), and testing sometimes requires server changes. Supporting types matching

provides a way to introduce new or private ecosystems without rewriting schema:

```
<idServer prefix="mvah">https://markspersonaltranslationproject.fr</idServer>
...
<systemId type="x-mvah">
   <id>idInMyPersonalFormat</id>
   <myDetail>something-that-interests-me</myDetail>
</systemId>
```

The type of all x-* systemIds should correspond to an idServer declaration.

3.2.2 ISSUES TO CONSIDER FOR 0.2

None.

3.3 canonSpec

3.3.1 PROPOSED CHANGES FOR 0.1

3.3.1.1 Remove the .*2 components

These variants of three components correspond to longstanding inconsistencies in the Canons.xml file, caused by inconsistent use of DAN/DAG and EST/ESG in canons of different scope for the same tradition (eg the OT part of the Armenian Bible canon does not match the Armenian OT canon). Also, there is no JER in the Greek Orthodox canon.

3.3.1.2 Remove DAG and ESG

These Greek variants of DAN and EST are not used consistently, and make canon management harder (since, for any Catholic or Orthodox project, there are 4 possible permutations of DAN/DAG and EST/ESG). It seems preferable to get the structure of books from the versification file.

3.3.2 ISSUES TO CONSIDER FOR 0.2

3.3.2.1 Develop canonSpec

One day, canonSpecs should be able to use custom components, which begs the question of where and how those components would be defined. eg would we do everything inline within the metadata file (easy to create, hard to reuse) or would components be declared separately (easier to reuse, share between users, etc, but assumes an ecosystem).

CHAPTER

FOUR

RELATIONSHIPS

4.1 PROPOSED CHANGES FOR 0.1

None

4.2 ISSUES TO CONSIDER FOR 0.2

4.2.1 Expand @relationType enum

This mechanism could be used to represent other entry-to-entry relationships, eg between Bible text and related para-biblical material.

CHAPTER

FIVE

AGENCIES

5.1 PROPOSED CHANGES FOR 0.1

5.1.1 Allow multiple name elements, each with language attribute

See the identification section.

5.1.2 Less compulsory fields for upload variant

Some background... DBL Metadata is used as the basis of the job spec at the heart of uploading. In this variant more values are optional - notably @revision since this will be overwritten by the server in any case.

Right now, all the denormalized fields in the agencies section are required. However, the uids in the agencies section may only be changed via the DBL website, ie revisions attempting to change ownership will be rejected. This means that clients need to generate a lot of boilerplate, some of which has to be identical to the information on the server, and some of which is not validated for coherence by the server. So, eg, it is currently possible to change the name and url of a rightsHolder but not the uid, which is a recipe for utter confusion.

The proposal is to make all fields optional when revision > 1, to reduce boilerplate and to not create false expectations about what may be changed via the client.

5.1.3 contributor/content should be optional

This is the only role-type field that is required for contributors, and was probably a typo in the schema.

5.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

SIX

LICENSE

This is a new, optional section describing any license under which the burrito has been distributed.

6.1 PROPOSED CHANGES FOR 0.1

(Needs to handle bilateral licenses, “public” licenses and public domain. Put things we can do now without scaring IPR people here.)

6.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

SEVEN

TYPE

7.1 PROPOSED CHANGES FOR 0.1

7.1.1 REAP-compatible isConfidential

Apparently REAP uses more than two states to represent the degree of confidentiality of a project. It would make sense for DBL to use the same system. It will be quite important to make sure that migration does not make previously confidential projects visible.

7.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

EIGHT

FORMAT - OVERVIEW

The format section is highly medium-dependent. The details for each of the five existing media are provided in subsequent chapters.

8.1 PROPOSED CHANGES FOR 0.1

8.1.1 Conventions

Conventions are intended to provide a mechanism for subtyping media, allowing greater flexibility along with improved server-side checking and bundle-consumer visibility.

```
<convention type="structure" version="1.0">usx-dirs</convention>
```

Zero or more convention elements would be allowed for each entry. In the absence of a convention, the entry may or not comply, ie *caveat emptor*.

@type is one of

- **structure** ie the "directory" structure, eg "usx-dirs"
- **content-format** ie the standard to which specific resources comply, eg "tagged-text" for print
- **content** ie the actual content of resources, eg, "usx-refs"

@version is required since conventions are likely to evolve.

The enum of conventions will need to be defined in consultation with stakeholders. x-.* should be supported for emerging conventions.

8.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

NINE

TEXT FORMAT

9.1 PROPOSED CHANGES FOR 0.1

9.1.1 Remove versedParagraphs

This can be replaced with a convention:

```
<convention type="structure" version="1.0">versed-paras</convention>
```

9.1.2 Syntax

This provides the USFM and USX version.

```
<usfmVersion>3.0</usfmVersion>
<usxVersion>3.0</usxVersion>
```

9.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

TEN

AUDIO FORMAT

10.1 PROPOSED CHANGES FOR 0.1

10.1.1 Conventions

- whole-chapter
- book-dirs
- wav-sources

10.1.2 Roles

- book-introduction
- audio-timing

10.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

ELEVEN

VIDEO FORMAT

11.1 PROPOSED CHANGES FOR 0.1

11.1.1 Conventions

- whole-chapter
- book-dirs
- roles-in-uris (a Nathanael wizard convention for encoding roles in filenames)

11.1.2 Roles for non-canonical video files

It may be possible to reuse some USFM peripheral “roles”, and the list will probably need to be extended after consultation with sign language stakeholders.

- bible-menu
- book-menu
- frontmatter
- backmatter
- copyright
- book-introduction

11.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

TWELVE

PRINT FORMAT

12.1 PROPOSED CHANGES FOR 0.1

12.1.1 Add print-oriented roles

- printBody
- printCover
- printThumbnail

12.1.2 Metadata for thumbnail JPEG

- width
- height
- colorModel

12.1.3 fonts element should be optional and never empty

This is a schema error, ie it should be

```
element fonts { printFormatFontElement+ }?
```

not

```
element fonts { printFormatFontElement* }
```

12.1.4 Support Biblica's Tagged Text Toolbox format

This variant of print content is essentially InDesign XML.

```
<convention type="structure" version="1.0">tagged-text</convention>
```

We need to rework the metadata fields too.

12.1.5 Enforce exactly one publication

Print entries are always an expression and it is hard to imagine a multiple-publication scenario.

12.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

THIRTEEN

BRAILLE FORMAT

13.1 PROPOSED CHANGES FOR 0.1

13.1.1 liblouis => brailleConvertor

LibLouis is **almost** ubiquitous for braille transcription in 2019, but other options do exist.

<brailleConvertor>liblouis-3.7.1</brailleConvertor>

13.1.2 table/source => table/src

This is just for consistency.

13.1.3 Enforce exactly one publication

Braille entries are always an expression and it is hard to imagine a multiple-publication scenario.

13.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

FOURTEEN

LANGUAGES

14.1 PROPOSED CHANGES FOR 0.1

14.1.1 BCP 47

This is the currently preferred way to store language information, as it can include most other standards and can handle minority languages and dialects that are unlikely to be supported by the older standards such as ISO 639-3.

It seems that the field currently called "ldml" is actually closer to BCP 47. This should be renamed "bcp47" and we ned a regex that handles all BCP 47 permutations.

14.1.2 Remove fields incorporated into BCP 47

- ISO
- scriptCode
- script
- numerals

14.1.3 Multiple Language Support

In the current schema there is exactly one language element. To support multiple languages we should adopt a structure similar to the current countries structure, ie a "languages" wrapper with one or more "language" child element. Exactly one of those languages should be marked as the default:

```
<languages>
    <language lang="fr" default="true">
    ...
    </language>
</languages>
```

The name and nameLocal fields need rethinking to support multiple languages, as per the identification section.

14.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

FIFTEEN

COUNTRIES

15.1 PROPOSED CHANGES FOR 0.1

15.1.1 Multiple Language Support

See the identification section.

15.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

SIXTEEN

NAMES

16.1 PROPOSED CHANGES FOR 0.1

Multiple abbr, short and long elements, distinguished by language attribute.

16.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

SEVENTEEN

MANIFEST

17.1 PROPOSED CHANGES FOR 0.1

17.1.1 Drop progress attribute

17.1.2 Tighten checksum regex

The regex currently allows S3 part suffixes which should never have been present in any metadata.

17.1.3 Drop containers

ie the manifest should always be a flat list of resources with fully-qualified uris.

17.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

EIGHTEEN

PUBLICATIONS

18.1 PROPOSED CHANGES FOR 0.1

18.1.1 Drop scope

18.1.2 Multiple Language Support

See the identification section.

18.1.3 Drop scope

18.1.4 metaContent

This would allow content elements to have child elements for supporting content. The first concrete use case is for timing files, which are closely related to audio or video files, but which appear as separate entries within the manifest:

```
<content src="MAT.usx" name="book-mat" role="MAT">
    <metaContent src="timing/MAT.xml"/>
</content>
<content src="MRK.usx" name="book-mrk" role="MRK">
    <metaContent src="timing/MRK_1-6.xml" role="MRK 1-6"/>
    <metaContent src="timing/MRK_7-16.xml" role="MRK 7-16"/>
</content>
```

18.1.5 Peripherals ids in metadata role enum

This enables the tagging of extra-canonical content without relying on well-known file names. The list, from the USFM 3 spec, would be

- **abbreviations**: Table of abbreviations
- **alphacontents**: Alphanumeric Contents
- **chron**: Chronology
- **cnc**: Concordance
- **contents**: Table of Contents
- **cover**: Cover
- **foreword**: Foreword

- **glo**: Glossary
- **halftitle**: Half Title Page
- **imprimatur**: Imprimatur
- **intbible**: Introduction to the Bible
- **intdc**: Deuterocanon Introduction
- **intepistles**: Introduction to Epistles
- **intgospels**: Introduction to Gospels
- **inthist**: Introduction to History
- **intnt**: Introduction to New Testament
- **intot**: Introduction to the Old Testament
- **intpent**: Introduction to the Pentateuch
- **intpoetry**: Introduction to Poetry
- **intprophesy**: Introduction to Prophecy
- **lxxquotes**: Quotes from LXX in NT
- **maps**: Map Index
- **measures**: Weights and Measures
- **ndx**: Names Index
- **preface**: Preface
- **promo**: Promotional Page
- **pubdata**: Publication Data
- **spine**: Spine
- **tdx**: Topical Index
- **title**: Title Page

18.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

NINETEEN

SOURCE

19.1 PROPOSED CHANGES FOR 0.1

19.1.1 Drop @name

19.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

TWENTY

COPYRIGHT

20.1 PROPOSED CHANGES FOR 0.1

20.1.1 Language attribute for statementContent

To support multiple languages, we would need to add a language attribute to each statement. This could be optional if there is only one language in the languages section.

20.2 ISSUES TO CONSIDER FOR 0.2

None

CHAPTER

TWENTYONE

PROMOTION

21.1 PROPOSED CHANGES FOR 0.1

21.1.1 Replace promoVersionInfo with statementContent

Right now the promotion section is similar to but confusingly different to the copyright section. A more coherent structure that also allows plain text promotional material would be

/DBLMetadata/promotion (Exactly 1)

- **statementContent[@type='xhtml']/*** (0 or 1 xml)
  - Promotional material in DBL's subset of XHTML (must be valid XML, ie tags must match.)
  - schema: dbl/2_2/dbl-xhtml
- **statementContent[@type='plain']** (0 or 1 string)
  - Promotional material in plain text

21.2 ISSUES TO CONSIDER FOR 0.2

None.

CHAPTER

TWENTYTWO

PROGRESS

22.1 PROPOSED CHANGES FOR 0.1

This section should be removed since it has never been used and does not contain useful data.

22.2 ISSUES TO CONSIDER FOR 0.2

None.