BookReporter.py

The BookReporter python script is a *prototype* developer tool that wraps the ePubCheck java application. The script has one primary purpose:

- Demonstrate of the value and uses that can be made through programmatic processing of check results rendered in JavaScript Object Notation (.json)

To demonstrate this value, the prototype offers features that allow its use as a tool for:

- Content creators to address ePub development scenarios
- Content managers for use in content management scenarios to verify ePub quality for relatively large volumes of books

ePub Development Scenarios

- During iterative refinement of an ePub, provide feedback on both any ePub errors found in given book and on changes in the book over time.
- Provide a relatively easy to consume "inventory" of the ePub's contents and the usage of ePub features in the publication.

Example:

Working in a directory with a single ePub, the first time a check is run, consider the following example output:

```
z:\work\TestData\BNePub.pp\NBRTest>BookReporter -w -u
BookReporter Tool: Check ePubs and optionally compare check results to a previous check
BookReporter is running on Python version: 2.7; no check time limit is being enforced. Some
books can take > 5 minutes to check...
--

Target Dir= .
--

File #1 (of 1) : .\9780307272300_epub.v8.epub has 3 severe errors (0 FATAL and 3 ERROR
messages)

    Messages Summary:
    ERROR messages:
      1: OPF-023: Expected 1 metadata definition tag but found '0'. (1 occurrence)
      2: HTM-020: Content file doesn't contain xml:lang attribute. (43 occurrences)
      3: OPF-032: Guide references 'OEBPS/images/Lars_9780307272300_epub_cvt_r1.jpg' which is
not a valid 'OPS Content Document'. (1 occurrence)
    --
    WARNING messages:
      1: HTM-016: HTML5 DOCTYPE definition within ePub v2. (1 occurrence)
      2: CSS-014: CSS Selector font-size attribute 'font-size' is declared using relative size
(eg, Percentage, keyword, or em multiple). (69 occurrences)
      3: CSS-010: Content file contains at least one inline style declaration. (1 occurrence)
      4: HTM-021: Content file doesn't contain lang attribute. (43 occurrences)
      5: OPF-003: Item 'iTunesMetadata.plist' exists in the ePub file, but is not declared in
the OPF manifest. (1 occurrence)
    --
    USAGE messages:
      1: CSS-012: Document contains 2 CSS files. (41 occurrences)
      2: HTM-005: An external reference was found. (1 occurrence)
    --
--
```

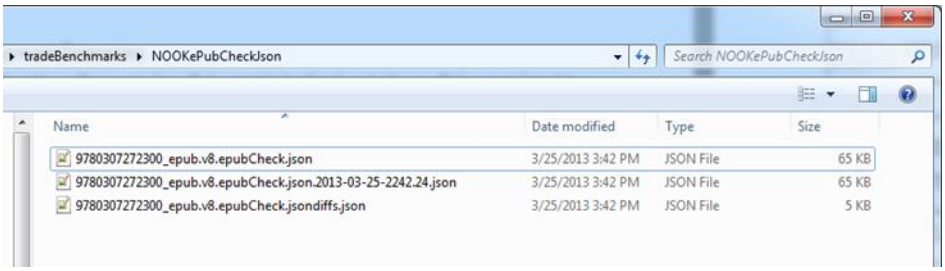
In this example, there are small number of errors that are likely to adversely affect the presentation of the content, 5 warning messages, describing problems that should at least be investigated, and 2 "Usage" messages that document use of particular ePub features. After a series of changes, a subsequent run of BookReporter produced the following output:

```
z:\work\TestNBR>BookReporter
BookReporter Tool: Check ePubs and optionally compare check results to a previous check
BookReporter is running on Python version: 2.7; no check time limit is being enforced. Some books can take >
5 minutes to check...
--

Target Dir= .
--
File #1 (of 1) : .\9780307272300_epub.v8.epub has 3 severe errors (0 FATAL and 3 ERROR messages)
    Found an older, differing, json output file for .\9780307272300_epub.v8.epub; saving the older version
as: '.\ePubCheckJson\9780307272300_epub.v8.epubCheck.json.2013-03-26-2358.07.json'

    ePubCheck results comparison
    --
    New file: z:\work\TestNBR\.\9780307272300_epub.v8.epub checked on: 03-26-2013 16:58:38
    Old file: z:\work\TestNBR\.\9780307272300_epub.v8.epub checked on: 03-25-2013 13:58:48
    --
    Summary: publication metadata changes: 2
             manifest item changes: 5
    --
    Publication property changes:
      Properties changed:
        'isScripted'changed --  new value: 'False'; old value: 'True'
        'title'changed --  new value: 'The Girl who Played with Fire'; old value: 'The Girl who Once Played
with Fire'
    --
    Publication manifest item changes:
      Manifest items added:
        'c02'

      Manifest items removed:
        'c07Duplicate'

      Manifest item property changes:
        Manifest item ID: 'toc' -- added reference to: 'OEBPS/Lars_9780307272300_epub_c31_r1.htm'
        Manifest item ID: 'toc' -- added reference to: 'OEBPS/Lars_9780307272300_epub_c06_r1.htm'
        Manifest item ID: 'toc' -- added reference to: 'OEBPS/Lars_9780307272300_epub_c32_r1.htm'
        Manifest item ID: 'toc' -- added reference to: 'OEBPS/Lars_9780307272300_epub_css_r1.css'
        Manifest item ID: 'toc' -- removed reference to: 'dummyReference'
        Manifest item ID: 'toc' -- property 'navigationOrder' changed -- newValue: 5; oldValue: 3
        Manifest item ID: 'adc' -- the associated file 'OEBPS/Lars_9780307272300_epub_adc_r1.htm' contents
changed
        Manifest item ID: 'tp' -- property 'navigationOrder' changed -- newValue: 3; oldValue: 5
        Manifest item ID: 'tp' -- property 'renditionLayout' changed -- newValue: reflowable; oldValue: fixed
        Manifest item ID: 'tp' -- property 'isFixedFormat' changed -- newValue: False; oldValue: True
    --

    Messages Summary:
    ERROR messages:
      1: OPF-023: Expected 1 metadata definition tag but found '0'. (1 occurrence)
      2: HTM-020: Content file doesn't contain xml:lang attribute. (43 occurrences)
      3: OPF-032: Guide references 'OEBPS/images/Lars_9780307272300_epub_cvt_r1.jpg' which is not a valid
'OPS Content Document'. (1 occurrence)
    --
    Generated messages from 9780307272300_epub.v8.epub are unchanged...
    --
--
```

In the sample output above (based on fabricated changes), you can see that two publication property changes were made and that one item was added and one was deleted from the ePub manifest, and that a number of manifest item properties were changed.

In addition to reporting results on the screen, you'll note that by default the results are saved in a subdirectory named "ePubCheckJson" in a .json form that can be readily consumed by other tools. The most recent check's results are saved in a file that matches the ePub file name with a file name suffix of ".epubCheck.json"; older versions of the check output file are also preserved, with the time they were created added to the file name as a timestamp. For example:

Note that a summary of the differences between two checks is also stored in a file, again named to match the book being checked, with a file name suffix of ".jsondiffs.json".



The BookReporter command line switch "-v" (for verbose mode output) causes the script to output additional information to the console, notably, the message output includes each location (file name, and when they are known, the line and column number) where the issue was detected that caused ePubCheck to output the message. As an example the output on the right shows the locations for several of the messages generated by the check of the text book used in this example. Note that when an issue is found more than 25 times in an ePub, the list is truncated and the total number of additional locations is shown.

```
Messages Summary:
 ERROR messages:
   1: OPF-023: Expected 1 metadata definition tag but found '0'. (1
occurrence)
         Lars_9780307272300_epub_opf_r1.opf

   2: HTM-020: Content file doesn't contain xml:lang attribute. (43
occurrences)
         OEBPS/Lars_9780307272300_epub_cvi_r1.htm
         OEBPS/Lars_9780307272300_epub_adc_r1.htm
                                        •
                                        •
                                        •
         OEBPS/Lars_9780307272300_epub_c17_r1.htm
         There are 18 additional locations for this message.

   3: OPF-032: Guide references
'OEBPS/images/Lars_9780307272300_epub_cvt_r1.jpg' which is not a valid 'OPS
Content Document'. (1 occurrence)
         Lars_9780307272300_epub_opf_r1.opf: line: 121 col: 126

 --
 WARNING messages:
   1: HTM-016: HTML5 DOCTYPE definition within ePub v2. (1 occurrence)
         OEBPS/Lars_9780307272300_epub_cvi_r1.htm

   2: CSS-014: CSS Selector font-size attribute 'font-size' is declared using
relative size (eg, Percentage, keyword, or em multiple). (69 occurrences)
         OEBPS/Lars_9780307272300_epub_css_r1.css: line: 5 col: 1
         OEBPS/Lars_9780307272300_epub_css_r1.css: line: 12 col: 1
         OEBPS/Lars_9780307272300_epub_css_r1.css: line: 18 col: 1
                                        •
                                        •
                                        •
         OEBPS/Lars_9780307272300_epub_css_r1.css: line: 163 col: 1
         OEBPS/Lars_9780307272300_epub_css_r1.css: line: 170 col: 1
         There are 44 additional locations for this message.
```

Content Management Scenarios

In some cases, ePubCheck is used as part of an ePub production process to assess the quality of several books at the same time. BookReporter.py supports these scenarios by enabling:

- Checking of all ePubs in a given directory
- Logging of check results to a central location
- Flexible output control for the .json files allowing comparison of book versions

Check all ePubs in a directory

In the example shown below, the –d command line option was used to test all of the ePubs in the "TestNBR" directory. (The output generated for books 2 through 5 has been removed for clarity.)

```
z:\work\TestData\BNePub.pp>bookreporter -d TestNBR -l
BookReporter Tool: Check ePubs and optionally compare check results to a previous check
Activity logging is being performed to z:\work\testLogs\BookReporter.TabDelimitedFile
BookReporter is running on Python version: 2.7; no check time limit is being enforced. Some books can take >
5 minutes to check...
--

Target Dir= TestNBR
--

File #1 (of 7) : TestNBR\9780307272300_epub.v8.epub has 3 severe errors (0 FATAL and 3 ERROR messages)

    Messages Summary:
    ERROR messages:
      1: OPF-023: Expected 1 metadata definition tag but found '0'. (1 occurrence)
      2: HTM-020: Content file doesn't contain xml:lang attribute. (43 occurrences)
      3: OPF-032: Guide references 'OEBPS/images/Lars_9780307272300_epub_cvt_r1.jpg' which is not a valid
'OPS Content Document'. (1 occurrence)
    --
--

File #2 (of 7) : TestNBR\9780307590626_ePub.v1.epub has 3 severe errors (0 FATAL and 3 ERROR messages)

    Messages Summary:
    ERROR messages:
      1: OPF-023: Expected 1 metadata definition tag but found '0'. (1 occurrence)
      2: HTM-020: Content file doesn't contain xml:lang attribute. (24 occurrences)
                                                    •
                                                    •
                                                    •
                                                    •
File #6 (of 7) : TestNBR\9780385534130_ePub.v1.epub has 3 severe errors (0 FATAL and 3 ERROR messages)

    Messages Summary:
    ERROR messages:
      1: OPF-023: Expected 1 metadata definition tag but found '0'. (1 occurrence)
      2: HTM-020: Content file doesn't contain xml:lang attribute. (53 occurrences)
      3: OPF-032: Guide references 'OEBPS/images/Gris_9780385534130_epub_cvt_r1.jpg' which is not a valid
'OPS Content Document'. (1 occurrence)
    --
--

File #7: ePubCheckJson is not an ePub, skipped…
```

Activity Logging

In the previous example, note the highlighted line which indicates that activity logging is being performed to the named tab delimited file. The simple log file includes a column header row and a single record for each check performed (the table is split below so it's content is legible):

logDate	logTime	logTool	PubDir	ePubFile	ePubPath	elpasedTi	checkTime	ePubVers	comparedTo
3/26/2013	14:36:11.352	NOOKBookReporter.py	TestNBR	9780307272300_epub.v8.epub	TestNBR\9780307272300_epub.v8.epub	5.739	5.671	2	z:\work\TestData\BNePub.pp
3/26/2013	14:36:22.541	NOOKBookReporter.py	TestNBR	9780307590626_ePub.v1.epub	TestNBR\9780307590626_ePub.v1.epub	11.185	11.113	2	z:\work\TestData\BNePub.pp
3/26/2013	14:37:36.706	NOOKBookReporter.py	TestNBR	9780316128568_ePub.v1.epub	TestNBR\9780316128568_ePub.v1.epub	8.359	8.151	2	NA
3/26/2013	14:37:42.616	NOOKBookReporter.py	TestNBR	9780385534130_ePub.v1.epub	TestNBR\9780385534130_ePub.v1.epub	5.909	5.682	2	NA
3/26/2013	14:38:51.247	NOOKBookReporter.py	.	9780307272300_epub.v8.epub	.\9780307272300_epub.v8.epub	5.206	4.959	2	z:\work\TestData\BNePub.pp
3/26/2013	14:40:38.376	NOOKBookReporter.py	TestNBR	9780307272300_epub.v8.epub	TestNBR\9780307272300_epub.v8.epub	5.59	5.572	2	NA

kTime	ePubVers	comparedTo	checkChanged	pubChanged	manifestChanged	messagesChanged	numFatal	numErrors	isScripted	hasFixedFormat
5.671	2	z:\work\TestData\BNePub.pp\TestNBR\9780307272300_epub.v8.epub	FALSE	Pub==	Mani==	Mess=	0	3	FALSE	FALSE
1.113	2	z:\work\TestData\BNePub.pp\TestNBR\9780307590626_ePub.v1.epub	FALSE	Pub==	Mani==	Mess=	0	3	FALSE	FALSE
8.151	2	NA	NA	NA	NA	NA	0	2	FALSE	FALSE
5.682	2	NA	NA	NA	NA	NA	0	3	FALSE	FALSE
4.959	2	z:\work\TestData\BNePub.pp\tradeBenchmarks\.\9780307272300_epub.v8.epub	TRUE	Pub-xA-xR-C2	Mani-A1-R1-C3	Mess=	0	3	FALSE	FALSE
5.572	2	NA	NA	NA	NA	NA	0	3	FALSE	FALSE

There are several things to note about this sample log:

- The "comparedTo" column names the file that was the source of the older check result
- The checkChanged, pubChanged, manifestChanged and messagesChanged columns indicate whether the check results changed between the newer and older check. In the first two rows the results didn't change. In row 5, the results did change (checkChanged = TRUE). The Publication metadata and the Manifest metadata changes are encoded in those columns, showing the count of Adds ("A"), Removed ("R"), and Changed ("C") properties. If there were no changes of that type, that is encoded as "xA", "xR" or "xC". This encoding allows the content manager to filter results and identify particular kinds of changes for closer review.
- The last four columns are self-explanatory.

Other features

BookReporter has a number of other features to support management of book collections, notably:

- --EanOnlyJsonNames: this switch is handy when book version numbers are encoded in the ePub file name. For example, consider two versions of the same book, 9781452100128_epub.v8.epub and 9781452100128_epub.v1.epub. The first boxed output example shows some of the check results for the older ePub version:

```
z:\work\VersionStudy\Books>BookReporter --EanOnlyJsonNames -d OldVersion -f 9781452100128_epub.v1.epub -j .\ePubCheckJson -l -
w -u
BookReporter Tool: Check ePubs and optionally compare check results to a previous check
Activity logging is being performed to z:\work\Webr2.0\epubrender\target\test-classes\webapp\cgi-
bin\testLogs\BookReporter.TabDelimitedFile
BookReporter is running on Python version: 2.7; no check time limit is being enforced. Some books can take > 5 minutes to
check...
--

Target Dir= OldVersion
--

File #1 (of 1) : OldVersion\9781452100128_epub.v1.epub has 2 severe errors (0 FATAL and 2 ERROR messages)

    Messages Summary:
    ERROR messages:
      1: OPF-023: Expected 1 metadata definition tag but found '0'. (1 occurrence)
                                                                    •
                                                                    •
                                                                    •
```

When --EanOnlyJsonNames is used, the resulting .json files are named using only the 13 digit Ean name; all file name characters after the 13 digits are omitted. In this case, the .json file holding the check results is named `9781452100128.ePubCheck.json`.

The example output on the following page (which has been snipped for clarity) shows the results when the newer version of the ePub is checked and the results compared to the results of the check of the older ePub. Many manifest items had changed between the two ePub versions, as well as one ePub property.

```
z:\work\VersionStudy>BookReporter --EanOnlyJsonNames -f 9781452100128_epub.v8.epub -j .\ePubCheckJson -l -w -u
BookReporter Tool: Check ePubs and optionally compare check results to a previous check
Activity logging is being performed to z:\work\Webr2.0\epubrender\target\test-classes\webapp\cgi-
bin\testLogs\BookReporter.TabDelimitedFile
BookReporter is running on Python version: 2.7; no check time limit is being enforced. Some books can take > 5 minutes to
check...
--

Target Dir= .
--

File #1 (of 1) : .\9781452100128_epub.v8.epub has 2 severe errors (0 FATAL and 2 ERROR messages)
    Found an older, differing, json output file for .\9781452100128_epub.v8.epub; saving the older version as:
'.\ePubCheckJson\9781452100128.ePubCheck.json.2013-03-26-2223.40.json'

    ePubCheck results comparison
    --
    New file: z:\work\VersionStudy\Books\.\9781452100128_epub.v8.epub checked on: 03-26-2013 15:24:01
    Old file: z:\work\VersionStudy\Books\OldVersion\9781452100128_epub.v1.epub checked on: 03-26-2013 15:23:35
    --

    Summary: publication metadata changes: 1
             manifest item changes: 69
    --

    Publication property changes:
      Properties changed:
        'creator'changed --  new value: '[u'Bakerella']'; old value: '[u'Bakeiella']'
    --

    Publication manifest item changes:

      Manifest item property changes:
        Manifest item ID: 'ch37' -- the associated file 'OPS/045-chapter37.html' contents changed
        Manifest item ID: 'ch37' -- property 'compressedSize' changed -- newValue: 5389; oldValue: 5489
        Manifest item ID: 'ch37' -- property 'uncompressedSize' changed -- newValue: 5389; oldValue: 5489
        Manifest item ID: 'ch37' -- the associated file 'OPS/045-chapter37.html' contents changed
        Manifest item ID: 'ch37' -- property 'compressedSize' changed -- newValue: 5389; oldValue: 5489
        Manifest item ID: 'ch37' -- property 'uncompressedSize' changed -- newValue: 5389; oldValue: 5489
        Manifest item ID: 'titlepage' -- the associated file 'OPS/002-titlepage.html' contents changed
        Manifest item ID: 'titlepage' -- property 'compressedSize' changed -- newValue: 533; oldValue: 745
        Manifest item ID: 'titlepage' -- property 'uncompressedSize' changed -- newValue: 533; oldValue: 745
        Manifest item ID: 'img125' -- the associated file 'OPS/images/logo.jpg' contents changed
        Manifest item ID: 'img125' -- property 'compressedSize' changed -- newValue: 27563; oldValue: 7373
        Manifest item ID: 'img125' -- property 'uncompressedSize' changed -- newValue: 27563; oldValue: 7373
        Manifest item ID: 'ch41' -- the associated file 'OPS/049-chapter41.html' contents changed
        Manifest item ID: 'ch41' -- property 'compressedSize' changed -- newValue: 4545; oldValue: 4612
        Manifest item ID: 'ch41' -- property 'uncompressedSize' changed -- newValue: 4545; oldValue: 4612
        Manifest item ID: 'ch40' -- the associated file 'OPS/048-chapter40.html' contents changed
        Manifest item ID: 'ch40' -- property 'compressedSize' changed -- newValue: 7993; oldValue: 8116
        Manifest item ID: 'ch40' -- property 'uncompressedSize' changed -- newValue: 7993; oldValue: 8116
        Manifest item ID: 'ch43' -- the associated file 'OPS/051-chapter43.html' contents changed
        Manifest item ID: 'ch43' -- property 'compressedSize' changed -- newValue: 4429; oldValue: 4509
        Manifest item ID: 'ch43' -- property 'uncompressedSize' changed -- newValue: 4429; oldValue: 4509
        Manifest item ID: 'ch42' -- the associated file 'OPS/050-chapter42.html' contents changed
        Manifest item ID: 'ch42' -- property 'compressedSize' changed -- newValue: 8823; oldValue: 8957
        Manifest item ID: 'ch42' -- property 'uncompressedSize' changed -- newValue: 8823; oldValue: 8957
        Manifest item ID: 'ch45' -- the associated file 'OPS/053-chapter45.html' contents changed
        Manifest item ID: 'ch45' -- property 'compressedSize' changed -- newValue: 3548; oldValue: 3595
        Manifest item ID: 'ch45' -- property 'uncompressedSize' changed -- newValue: 3548; oldValue: 3595
                                                                  •
                                                                  •
                                                                  •
        Manifest item ID: 'ch20' -- property 'compressedSize' changed -- newValue: 3224; oldValue: 3152
        Manifest item ID: 'ch20' -- property 'uncompressedSize' changed -- newValue: 3224; oldValue: 3152
        Manifest item ID: 'ch61' -- the associated file 'OPS/071-chapter61.html' contents changed
```

BookReporter.py Prerequisites

BookReporter requires either Python v 2.7.x or 3.3.x to run. The script and the ePubCheck.jar file must exist in the same directory, or the --applicationJar command line argument must be used to specify the jar file's location. It is handy if you put the script and .jar directory on the path, and associate .py with the Python interpreter, but these steps are not required.

BookReporter Command Line Help

The BookReporter tool has command line reference available. Use the "-h" option to display it:

```
BookReporter Tool: Check ePubs and optionally compare check results to a previous check
Usage: BookReporter.py [OPTION]
BookReporter: ePubCheck all ePub files in the target directory, potentially preserving
                generated .JSON output files, compare results to prior checks if old results are found.

Options:
  -h, --help            show this help message and exit
  -d TARGET, --directory=TARGET
                        Directory on which ePubCheck will be run, default
                        is the current working directory
  -f TARGETFILE, --file=TARGETFILE
                        File or comma separated list of files to run the check
                        on; if -f is omitted, all files in the target
                        directory will be checked
  --NoSaveJson          Do NOT save ePubcheck .json output files
  --NoCompareJson       Do NOT compare the json created during this check with
                        the most recently saved .json result, if found.
  --EanOnlyJsonNames    Use this flag to force .json file names to use EAN-
                        only naming convention, <ean>.ePubCheck.json, not
                        <file_name>.ePubCheck.json names. Files not conforming
                        to EAN-first naming pattern will use the <file_name>
                        convention
  -j JSONDIR, --jsonDir=JSONDIR
                        if the -s switch is used, ePubCheck .json output files
                        will be preserved in either the location specified by
                        the -e switch, or if -e is omitted, stored in
                        <targetDir>\ePubCheckJson
  -v, --verbose         Show all messages grouped by type
  -q, --Hide_errors     'Quiet' output mode; don't list FATAL and ERROR
                        messages; by default these errors are always displayed
                        on the console
  -w, --warning         Show WARNING messages; by default, these messages are
                        not shown on the console
  -u, --usage           Show USAGE messages, by default, these messages are
                        not shown on the console
  -l, --logging         Enable logging to a tab-delimited file
  --logdir=LOGDIR       Log file location used by this tool, defaults to the
                        value of the environment variable "EPUBCHECK-LOGS";  if
                        EPUBCHECK-LOGS is defined and a valid directory, logging is
                        enabled to that directory automatically (-l is not
                        required if EPUBCHECK-LOGS is defined) if EPUBCHECK-LOGS is
                        undefined and logging is enabled, logs are written in
                        the current working directory. If EPUBCHECK-LOGS is
                        defined, automatic logging can be disabled by using
                        the "--logdir none" switch.
  --logfile=LOGFILE     Log file name used by this tool,
                        default=BookReporter.TabDelimitedFile
  --applicationJar=APPJAR
                        if specified,the named jar will be used; if not
                        specified, Z:\work\ePubCheck\com.adobe.epubcheck\targe
                        t\epubcheck-4.0.0-SNAPSHOT.jar in this script's
                        directory will be used
  --jarArgs=JARARGS     Any args specified with the --jarArgs switch will be
                        passed to the applicationJar (ePubCheck-4.0.0-SNAPSHOY); by
                        default, no jarArgs are passed
  --timeout=TIMEOUTVAL  Abort a ePubCheck process that takes longer than
                        the --timeout nnn in seconds. NOTE: This setting is
                        ignored unless you are using Python 3.3 or later
```