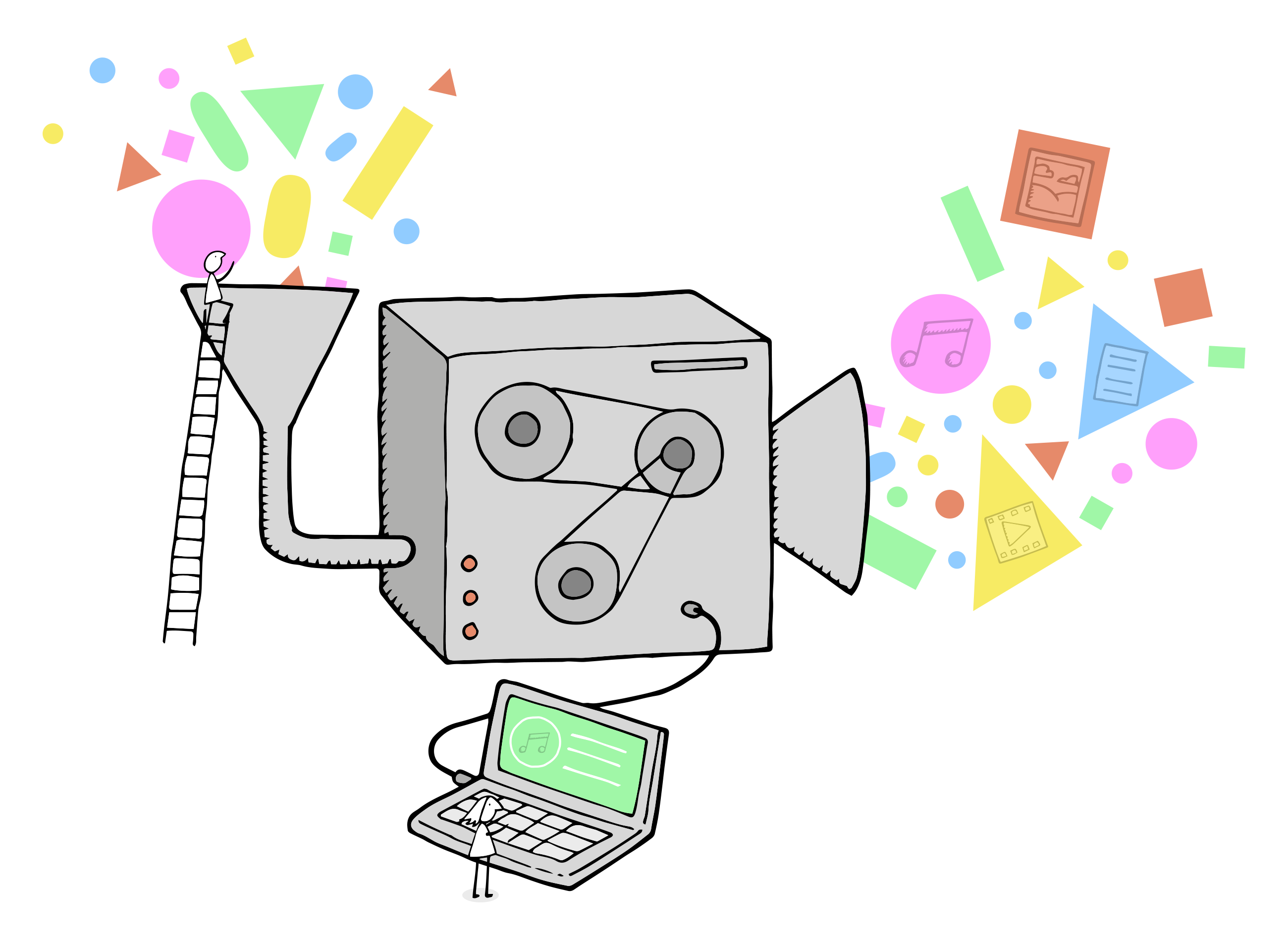
**File formats: Characterization and validation**

**Exercise Booklet**

**Bodleian Libraries**

**31/01/2018**

**Version 1.0**

****

*Digital Preservation Business Case Toolkit* [*http://wiki.dpconline.org/*](http://wiki.dpconline.org/)*, CC-BY-NC 3.0*

**Exercise 1: File format signatures & PRONOM**

**Goal of activity:** to understand how to investigate files in a hex editor, especially how to identify file format signatures, and where to find textual information and embedded metadata.

**Folder of test files:** 01\_HexEdit-PRONOM

|  |
| --- |
| 1. 01\_Toby-the-sapient-pig.tif   File type: TIFF  PUID: fmt/353  **Questions:**   * Identify the file format signature using PRONOM * Is it big-endian or little-endian? * What is the ICC Profile (also known as the colour space)? * Can you find out what this TIFF file was converted from? (Hint: it is in the XMP metadata in the header) * What was the created date? |
| 1. 02\_divinityschool\_ElizabethI.jpg   File type: JPEG  PUID: fmt/44  **Questions:**   * Identify the file format signature using PRONOM * Can you find the embedded metadata? * What software created this jpeg? * What was the created date? |
| 1. 03\_sound.wav   File type: Waveform  PUID: fmt/141  **Questions:**   * Identify the file format signature using PRONOM * Can you find out what the sound is supposed to be? (Hint: search the end of the file) |
| 1. ***OPTIONAL:*** 04\_Table2   File type: WordPerfect 5.0  PUID: x-fmt/393  **Questions:**   * Identify the file format signature using PRONOM |
| 1. ***OPTIONAL:*** 05\_Word-file-with-image   File type: .doc (word document)  PUID: fmt/40  **Questions/Activities:**   * Identify the file format signature using PRONOM * Can you find some of the text from the word document? Write down a sentence here. * Some files are made of multiple bitstreams, like this one. It contains an embedded image file. Find the file format signature. What is it? Verify it again the signature in PRONOM. (Hint: the file format signature has been included in the packet) * What was the Last Modified date of the embedded file? |
| 1. 06\_Crossing-the-Bar   File type: .rtf (rich text file)  PUID: fmt/355  **Questions:**   * Identify the file format signature using PRONOM |

**Appendix to Exercise 1: PRONOM signatures**

Below are the file signatures for each of the file format types found in Exercise 1 for your reference. You may also look them up on the PRONOM website: <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

****

****

****

****

****

****

****

**Exercise 2: File identification in DROID**

**Goal:** To run a DROID report over a number of files and practise reading the results to understand file format identification and what it can find and cannot find.

**Folder:** 02\_DROID

|  |
| --- |
| 1. Run a DROID report over the folder 02\_DROID. |
| 1. Answer the following questions: |
| * 1. How many files are showing a warning sign next to the extension? |
| * 1. Can you explain what each warning sign means for each file? |
| * 1. How many different PDF versions are there? |
| * 1. What file formats are identified by *Container* method? |
| * 1. Why is the *[insert file name].txt* and *[insert file name].csv* files are identified by the *Extension* method? (hint: check PRONOM) |
| * 1. How many zip folders are there? List what is inside them and make any notes about interesting findings. |
| * 1. Are there any unidentified files? If so, what one(s)? What would you do about the file(s) going forward? |
| * 1. Pick a file format from the list that you might not be familiar with and investigate it in PRONOM by the PUID. Note anything interested about it. |
| * 1. Did you have any identifications that had more than one option? If so, what one(s) did you find? |

**Exercise 3: Validation with JHOVE**

**Goal:** To run validation in JHOVE and look at the characterization outputs in the JHOVE report.

**Folder:** 03\_JHOVE

**PART 1.** Run JHOVE validation over the following files (**Note**: Be sure to select the correct module before running the report):

|  |
| --- |
| 1. **JP2** – 01\_balloon\_trunc1.jp2   Does the file validate? Y or N  If it does not validate, what is the error message? |
| 1. **TIFF** – 02\_CatTIFF\_01.tif   Does the file validate? Y or N  If it does not validate, what is the error message?  What version TIFF is it?  What is the Byte Order?  What is the compression scheme?  What is the image width and length? |
| 1. **XML** – 03\_Metadata-3.xml   Does the file validate? Y or N  If it does not validate, what is the error message?  What metadata schema is it? |
| 1. **AIFF** – 04\_wind-sound-0595.aiff   Does the file validate? Y or N  If it does not validate, what is the error message? |
| 1. ***OPTIONAL:* JPEG** -– 05\_First\_View\_of\_Earth\_from\_Moon.jpg   Does the file validate? Y or N  If it does not validate, what is the error message?  What is the compression type?  What is the Byte Order?  What is the image width and length? |
| 1. ***OPTIONAL:* PDF** – 06\_Digital-Preservation-Policy-Manual-NZ.pdf   Does the file validate? Y or N  If it does not validate, what is the error message?  What version of PDF is it?  What was the creation date? |

**PART 2.** Using HexEdit and PRONOM, edit the **file signature** of the following files and run a JHOVE validation (Note: be sure to save them as NEW FILES)

|  |
| --- |
| 1. **TIFF** – 02\_CatTIFF\_01.tif   Does the file validate? Y or N |
| 1. J**PEG** – 05\_First\_View\_of\_Earth\_from\_Moon.jpg   Does the file validate? Y or N |

**PART 3. *OPTIONAL*** Using HexEdit, edit the **BODY of the file** (Note: be sure NOT to touch any of the file signature, refer to PRONOM if necessary). Feel free to make as many changes to the document if you want. View it to see what changes have been made before running a JHOVE validation over them. (Note: Be sure to save them as NEW FILES. Use the original files to edit, so they will be valid)

|  |
| --- |
| 1. TIFF – 02\_CatTIFF\_01.tif   Does the file validate? Y or N  Open in an image viewer. What issues do you see with this? |
| 1. JPEG – 05\_First\_View\_of\_Earth\_from\_Moon.jpg   Does the file validate? Y or N  Open in an image viewer. What issues do you see with this? |

**Exercise 4: PDF/A Validation with VeraPDF**

**Goal:** to run validation over PDF/A files in VeraPDF and to examine the output report

**Folder:** 04\_VeraPDF

Run VeraPDF over the test set to see which ones validate and which ones do not. Make note of why validation fails. Have a look at why validation fails and see what you can understand about the complex PDF/A standard.

|  |
| --- |
| 1. **01\_TWG test suite A001-pdfa1-a.pdf**   Did the file validate? Y or N  Validation profile?  What was the failure reason? |
| 1. **02\_TWG test suite A001-pdfa1-b.pdf**   Did the file validate? Y or N  Validation profile?  What was the failure reason? |
| 1. **03\_TWG test suite A001-pdfa1-a.pdf**   Did the file validate? Y or N  Validation profile?  What was the failure reason? |
| 1. **04\_TWG test suite A002-pdfa2-a.pdf**   Did the file validate? Y or N  Validation profile?  What was the failure reason? |
| 1. ***OPTIONAL:* 05\_TWG test suite A027-pdfa3-e.pdf**   Did the file validate? Y or N  Validation profile?  What was the failure reason? |
| 1. ***OPTIONAL:* 06\_TWG test suite A027-pdfa3-d.pdf**   Did the file validate? Y or N  Validation profile?  What was the failure reason? |