The Bits In The Bytes

Understanding File Format Identification

|  |  |
| --- | --- |
| **Francesca Mackenzie** | **Andrea Hricikova** |
| *The National Archives, UK**United Kingdom**francesca.mackenzie@nationalarchives.gov.uk* | *The National Archives, UK**United Kingdom**andrea.hricikova@nationalarchives.gov.uk* |
| **Andrey Kotov** | **Kathryn Phelps** |
| *The National Archives, UK**United Kingdom**andrey.kotov@nationalarchives.gov.uk* | *The National Archives, UK**United Kingdom**Kathryn.phelps@nationalarchives.gov.uk* |

**Abstract – This workshop will provide practical experience of analyzing digital files to create signatures for format identification. It will build confidence in file-format analysis and develop participants’ understanding of a range of methods that can be applied to different files and content types. We will explore the approaches adopted by the major file format identification tools used by the digital preservation community. During this workshop, attendees will gain hands-on experience in the tools needed to contribute file format research to the open-source registry PRONOM; some participants will be analyzing the digits, or hex, of their files for potentially the first time. As well as being educational, file format identification is a lot of fun!**

**PRONOM as a tool ties in well with the themes of the conference. PRONOM is open source and used across the globe in the information management and digital preservation sectors, and beyond. It embodies the value of data for all and encourages understanding of file formats for future preservation needs.**

**PRONOM particularly embodies the key conference themes of community and exchange. We rely on so many talented file-format researchers around the world to analyze digital collections, flag issues and contribute to our shared knowledge of file formats. We want to continue the conversation with the digital preservation community and enable more people to participate in this collective endeavor.**

**Keywords – Collaboration, Hex, Community, Formats, Conversation.**

**Conference Topics – Exchange; Community**

# Overview

The workshop will be organized and run by the PRONOM file format team at The National Archives, UK. By the end of this workshop participants will be equipped with the knowledge and skills required to conduct file-format research and submit new entries to PRONOM. Attendees will gain an understanding of the theory behind creating file format signatures and will be able to apply this to their own digital collections. An understanding of what file formats you have in your repository is an essential part of digital preservation. Why? For the same reason it is essential for a conservator to know if a physical record is vellum or paper, digital preservation experts have to know what file formats are in their collection in order to effectively preserve them for future generations.

The workshop will be a mix of demonstration, lecture and hands-on activities. We would aim to make the workshop hybrids so that participants who cannot attend the conference in person would also be able to attend. Sample files will be provided but participants will be encouraged to bring their own unidentified file formats too.

We would additionally wish to elicit and capture feedback from participants on how they would prefer to continue engaging with and contributing to PRONOM in the future. How best can PRONOM facilitate file format research and collaborate with the community?

# Key Aims

The aim and benefits of the workshop is that participants should come away with:

*1) An understanding about file formats, signatures and the approaches available for characterizing files for digital preservation.*

*2) Learn how file format signatures are created, tested and submitted to PRONOM.*

*3) Have an increased knowledge of the range of (free) resources available to file format analysts and gain some practical experience using these tools.*

*4) An understanding of their own digital collections at a more binary level.*

*5) To feel confident enough to join the PRONOM community! Everyone can do file format research, and if they so wish, everyone should be able to contribute to research and help us improve.*

# Target Audience

Everyone is welcome who is willing to learn more about file formats and their identification. This workshop would be particularly useful for anyone who currently works or plans on working with the PRONOM technical registry, or wishes to develop a deeper understanding of format-based tools or workflows.

# Workshop Structure

Introduction to file format research (10 minutes)

File format signatures within PRONOM (20 minutes)

Activity: File format signatures using hex editors and file format specifications to create a signature. This will include reading byte streams, file format specifications and creating your own signature. (40 minutes)

Activity: Signature testing in which we will cover use of the skeleton suite, file format utility tool and testing your signature using DROID. (30 minutes)

BREAK (20 minutes)

Advanced signature development. This will cover container signature development and finding file samples. (40 minutes)

Activity: Open file format research and feedback (50 minutes)

This is a time for participants who want to stay to work on their own unidentified formats to stay and ask questions. It would also be valuable to ask the community how PRONOM can better support them.

# Other Information

### 4.1 Please Bring

Own laptop. It is also recommended to have an installation of DROID; installation of a text editor (e.g. notepad++ or sublime); installation of a hex editing software (e.g. HxD). Unidentified file formats also welcome.

## 4.2 Workshop Length

The workshop will be 3.5 hours and include a 20 minute break.

# REFERENCES

1. Manuscript Templates for Conference Proceedings, IEEE. http://www.ieee.org/conferences\_events/conferences/publishing/templates.html