# Implementing Metadata Standards

EaaSI Training Module #7

# During This Module

- What kind of metadata does EaaSI (the program of work) gather?
- What metadata has been directly incorporated into the EaaSI platform?
- Which standards are influencing EaaSI system design?
- How does EaaSI share/contribute its metadata to the field?


Image source: https://archive.org/download/Super_Cool_Shareware_PC-World_Digital_Concepts_1999/Super%20Cool%20Shareware%20%28PC-World%29%28Digital%20Concepts%29%281999%29.jpg

***But first...if someone asked, how would you describe a particular piece of software?***

# Categorizing Software

You could start by asking (and answering)...

- What is the software?
- Who is responsible for the software?
- How is the software distributed?
- What is the software made of?
- What is needed to run the software?

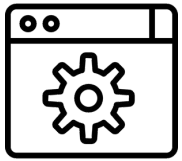
***NB:*** The questions above are taken from the Software Preservation Network's Software Metadata Recommended Format (SMRF) Guide!

| ***Property*** | Quick/Local Definition | ***Example*** |
|---|---|---|
| **Software Product Name** | Name of the software product as it appears on distribution mechanism or packaging | PrintMaster Gold Classic |
| **Software Product Version** | Version of the software product as it appears within system metadata (or on packaging if system metadata is unavailable) | 4.03.43 |
| **Distribution Mechanism** | Media carrier or file download by which the software is distributed to users | CD-ROM |
| **Installable Languages** | Installable languages for the Software Product (note: not for licenses) referenced within the software | English |
| **Executable Location** | Where the application is stored in the emulated computer's file system after installation. This must be a full path | C:\pmw\Pmw.exe |
| **Open File Format(s)** | File formats that can be opened by the software as represented within the software | Card Files (*.car); Banner Files (*.ban) |

Here is ***some*** example metadata that the EaaSI program of work gathers about software at the moment

# Every Metadata Field Has a Story

Every field in the prior table was selected because of its value to one or more aspects of our program of work, such as:

- Feature development for EaaSI Client 
- Search and discovery 
- Contributing to Wikidata 

- Automated dependency tracking and/or recommendations 

# EaaSI ≠ A Standard

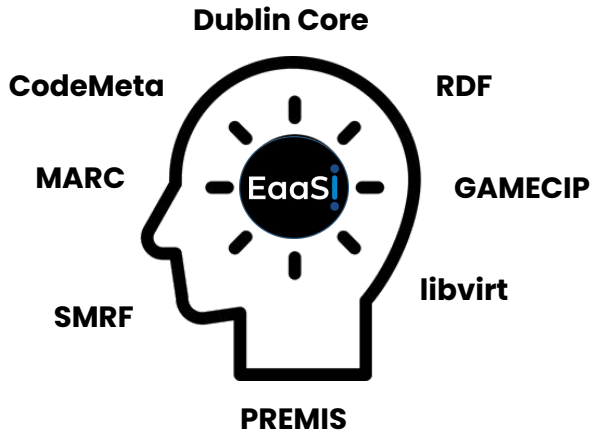

- EaaSI's metadata-related decisions are currently driven by:
  - Functional requirements of emulation, EaaS and/or the EaaSI UI
  - Grant deliverables (reporting benchmarks)
  - Perceived value to global digital preservation efforts in gathering information about legacy software
- EaaSI neither uses a singular centralized authority for software or emulation-relevant metadata - ***nor is trying to be one***
- Given funding and intended community, have focused on the needs of cultural heritage and research data management

# What *Is* a Standard, Then


- Metadata standards require consensus - or at least common understanding
- Structured requirements that ensure data can be properly interpreted across systems or individual platforms (like the EaaSI stack)
- Evolving demands of digital preservation are currently directing community energy towards creating metadata standards related to software preservation

# Looking for Inspiration

- EaaSI team wants the stack to draw on existing bibliographic and preservation standards…
- …but also make room for emerging efforts and expertise in standards for software preservation or related domains



# Mapping

- Where possible, EaaSI is making efforts to make crosswalks available between metadata captured by the team to existing standards
- Structural mapping now will make direct integration with repositories and catalogs easier down the road
- Occasionally difficult due to lack of emulation-specific properties in existing preservation and bibliographic standards
- EaaSI needs to describe software both in and of itself - a standalone “work” - and as part of a complex, emulated Environment

# Example Metadata Mapping

| ***EaaSI Metadata Gathering*** | ***MARC*** | ***MODS*** | ***Dublin Core*** | ***SMRF*** | ***CodeMeta*** | ***Wikidata QID*** |
|---|---|---|---|---|---|---|
| UUID (assigned by EaaS) | 024 1 - Other Standard Identifier, Universal Product Code https://www.loc.gov/marc/bibliographic/bd024.html<br><br>024 8 - Other Standard Identifier, Unspecified type of standard number or code https://www.loc.gov/marc/bibliographic/bd024.html | <identifier> https://www.loc.gov/standards/mods/userguide/identifier.html | Identifier http://purl.org/dc/terms/identifier | Identifier | identifier | unique identifier (Q6545185) https://www.wikidata.org/wiki/Q6545185 |
| Software Product Name | 245 - Title Statement https://www.loc.gov/marc/bibliographic/bd245.html<br><br>500 - General Note (to provide source of title) https://www.loc.gov/marc/bibliographic/bd500.html | <titleInfo> https://www.loc.gov/standards/mods/userguide/titleinfo.html | Title http://purl.org/dc/elements/1.1/title | Title | | title (Q783521) https://www.wikidata.org/wiki/Q783521 |
| Software Product Version | 250 - Edition Statement https://www.loc.gov/marc/bibliographic/bd250.html<br><br>251 - Version Information https://www.loc.gov/marc/bibliographic/bd251.html | <originInfo> <edition> https://www.loc.gov/standards/mods/userguide/originfo.html#edition | Relation http://purl.org/dc/terms/relation | Version | softwareVersion | software version (Q20826013) https://www.wikidata.org/wiki/Q20826013 |

# Example Metadata Mapping

| ***EaaSI Metadata Gathering*** | ***MARC*** | ***MODS*** | ***Dublin Core*** | ***SMRF*** | ***CodeMeta*** | ***Wikidata QID*** |
|---|---|---|---|---|---|---|
| Date Published | 260 - Publication, Distribution, etc. https://www.loc.gov/marc/bibliographic/bd260.html<br><br>542 - Information Relating to Copyright Status https://www.loc.gov/marc/bibliographic/bd542.html<br><br>388 - Time period of Creation https://www.loc.gov/marc/bibliographic/bd388.html | <originInfo><br><dateCreated> https://www.loc.gov/standards/mods/userguide/origininfo.html | Date http://purl.org/dc/elements/1.1/date<br><br>Date Created http://purl.org/dc/terms/created | Date, Date Subtype | datePublished | copyright date (Q59584702) https://www.wikidata.org/wiki/Q59584702<br><br>date of publication (Q1361758) https://www.wikidata.org/wiki/Q1361758 |
| Required Software (OS) | 753 $c - System Details Access to Computer Files; Operating System https://www.loc.gov/marc/bibliographic/bd753.html | <note> https://www.loc.gov/standards/mods/userguide/note.html | Relation http://purl.org/dc/elements/1.1/relation | Operating System | operatingSystem | operating system (Q9135) https://www.wikidata.org/wiki/Q9135 |
| Required Software (Other) | 538 - System Details Note https://www.loc.gov/marc/bibliographic/bd538.html | <note> https://www.loc.gov/standards/mods/userguide/note.html | Relation http://purl.org/dc/elements/1.1/relation | Additional Dependencies | | system requirements (Q2275513) https://www.wikidata.org/wiki/Q2275513 |

# Example Metadata Mapping

| ***EaaSI Metadata Gathering*** | ***MARC*** | ***MODS*** | ***Dublin Core*** | ***SMRF*** | ***CodeMeta*** | ***Wikidata QID*** |
|---|---|---|---|---|---|---|
| Selected Base Environment | 538 - System Details Note https://www.loc.gov/marc/bibliographic/bd538.html<br><br>753 - System Details Access to Computer Files https://www.loc.gov/marc/bibliographic/bd753.html | <note> https://www.loc.gov/standards/mods/userguide/note.html | Relation http://purl.org/dc/elements/1.1/relation | | softwareRequirements | software dependency (Q56859575) https://www.wikidata.org/wiki/Q56859575 |
| Default Save File Format | 256 - Computer File Characteristics https://www.loc.gov/marc/bibliographic/bd256.html<br><br>347 - Digital File Characteristics https://www.loc.gov/marc/bibliographic/bd347.html<br><br>516 - Type of Computer File or Data Note https://www.loc.gov/marc/bibliographic/bd516.html | <physicalDescription><form> https://www.loc.gov/standards/mods/userguide/physicaldescription.html#form<br><br><physicalDescription><internetMediaType> https://www.loc.gov/standards/mods/userguide/physicaldescription.html#internetmediatype<br><br><physicalDescription><digitalOrigin> https://www.loc.gov/standards/mods/userguide/physicaldescription.html#digitalorigin | Format http://purl.org/dc/elements/1.1/format | File Format | fileFormat | format (Q2085518) https://www.wikidata.org/wiki/Q2085518 |
| Open File Format(s) | 256 - Computer File Characteristics https://www.loc.gov/marc/bibliographic/bd256.html<br><br>347 - Digital File Characteristics https://www.loc.gov/marc/bibliographic/bd347.html<br><br>516 - Type of Computer File or Data Note https://www.loc.gov/marc/bibliographic/bd516.html | <physicalDescription><form> https://www.loc.gov/standards/mods/userguide/physicaldescription.html#form<br><br><physicalDescription><internetMediaType> https://www.loc.gov/standards/mods/userguide/physicaldescription.html#internetmediatype<br><br><physicalDescription><digitalOrigin> https://www.loc.gov/standards/mods/userguide/physicaldescription.html#digitalorigin | Format http://purl.org/dc/elements/1.1/format | File Format | fileFormat | format (Q2085518) https://www.wikidata.org/wiki/Q2085518 |

# Metadata is Labor

- Bulk of EaaSI-related metadata is currently gathered by a team of student employees at Yale University Library:
  https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure/#students
- Knowing what software metadata is - and how to find it - takes training and experience
- Automation efforts promise future improvements to efficiency, scale, and more time dedicated to creating new Environments rather than describing them

# Metadata Exchange

- Currently Environment and Software resource records can be passed between EaaSI nodes using OAI-PMH (see Module #6)
    - This *does not include* - yet! - much of the descriptive and structural metadata properties about software gathered by the student team
- Experimenting with serialization of expanded Environment and Software metadata via JSON and XML
- Relevant properties gathered by students are sent via spreadsheet to WikiDP

# Wikidata and WikiDP


- Wikidata for Digital Preservation (WikiDP) offers a custom portal for browsing and contributing to the Wikidata knowledge base in categories related to computing and software history
- Partnered with EaaSI to encourage conversion of our metadata into structured, linked open data
- Wikidata contributions are broadly reviewed and consensus-driven, so contributed metadata is purposefully not specific to the EaaSI stack
- Requires some additional evidence to bolster "notability" of new properties (i.e. file formats, software versions, and computing environments need to be formally differentiated)

# Software Preservation Network

- Affiliating with a relevant professional organization taps EaaSI into emerging software preservation process and standards

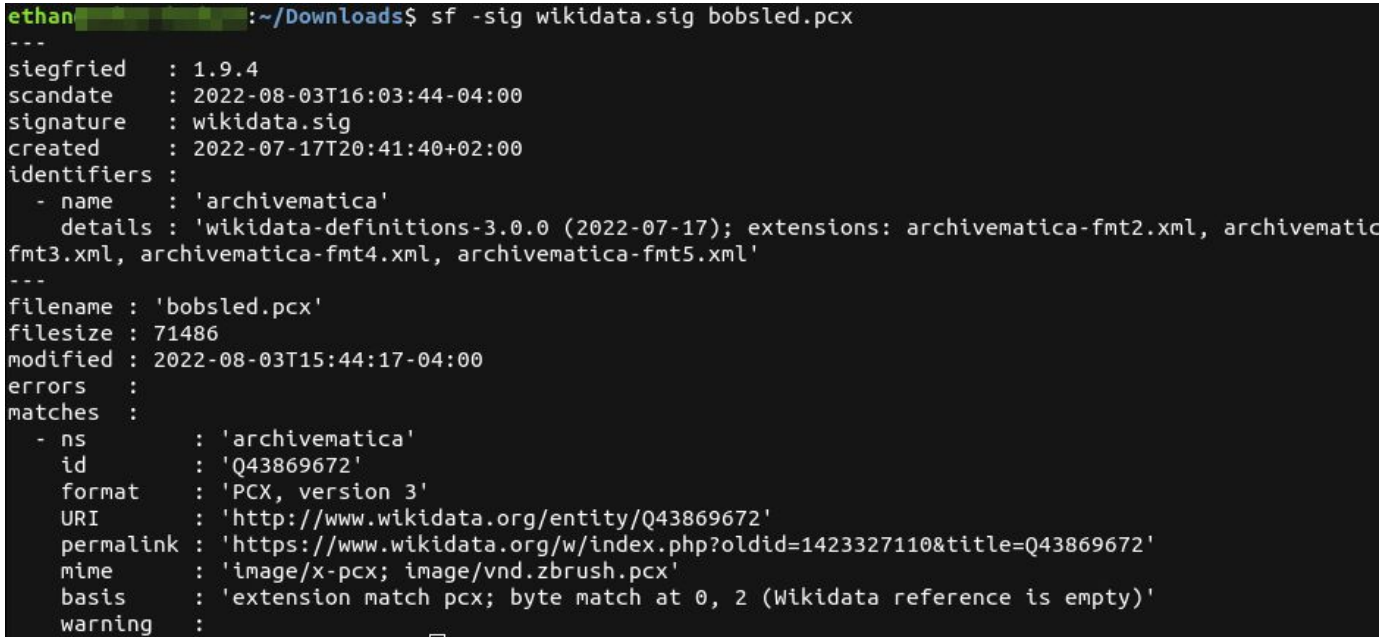
- Helps EaaSI to advocate for broader adoption and collection of metadata necessary for emulation

- Ensures collective, collaborative feedback on search behavior, future integrations, API design, mapping, etc.

# A Journey in Software Metadata via Emulation

*In this module we have been discussing software metadata for description, automation, discovery, etc.*

*How does all this actually come into play with emulation and the EaaSI platform? Let's walk through the process of emulating a file from scratch - charting *some* of the metadata (and relationships between pieces of metadata) necessary along the way.*



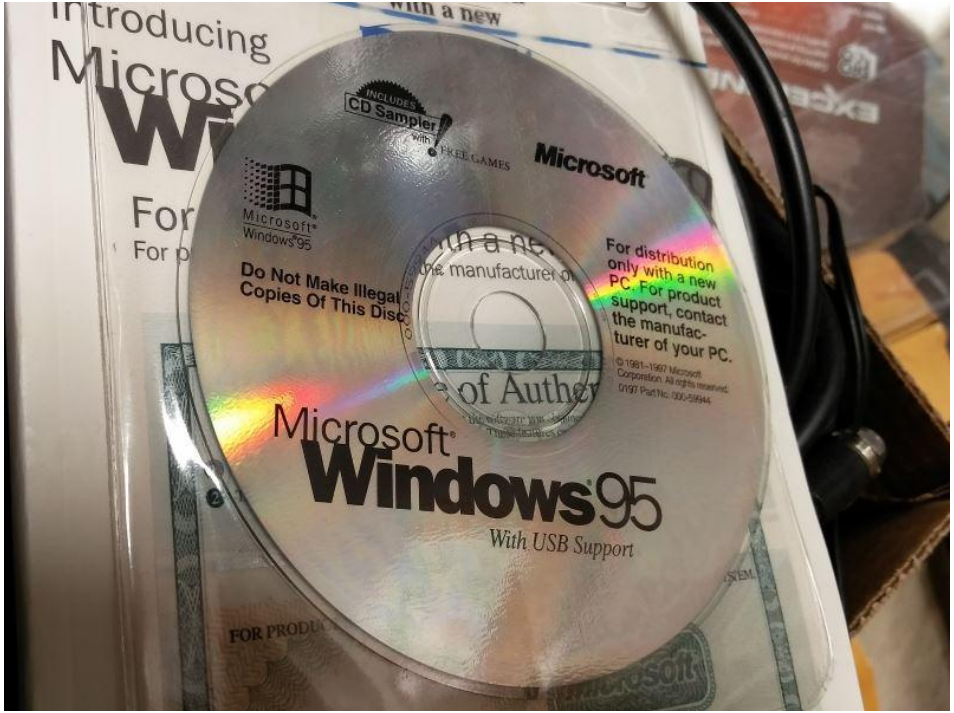

Step 1: I have a file - **“bobsled.pcx”** - that won’t open in contemporary software. Context clues - the extension name, file characterization - leads me to believe this was an image file made for/by a PC, probably sometime in the early to mid-1990s. We can guess it depicts a bobsled, but we can only know for sure by properly rendering it.

Step 2: I must choose a PC ***operating system*** to emulate: ***Windows***


Step 3: I must also choose a specific ***version*** to emulate: Windows ***95***



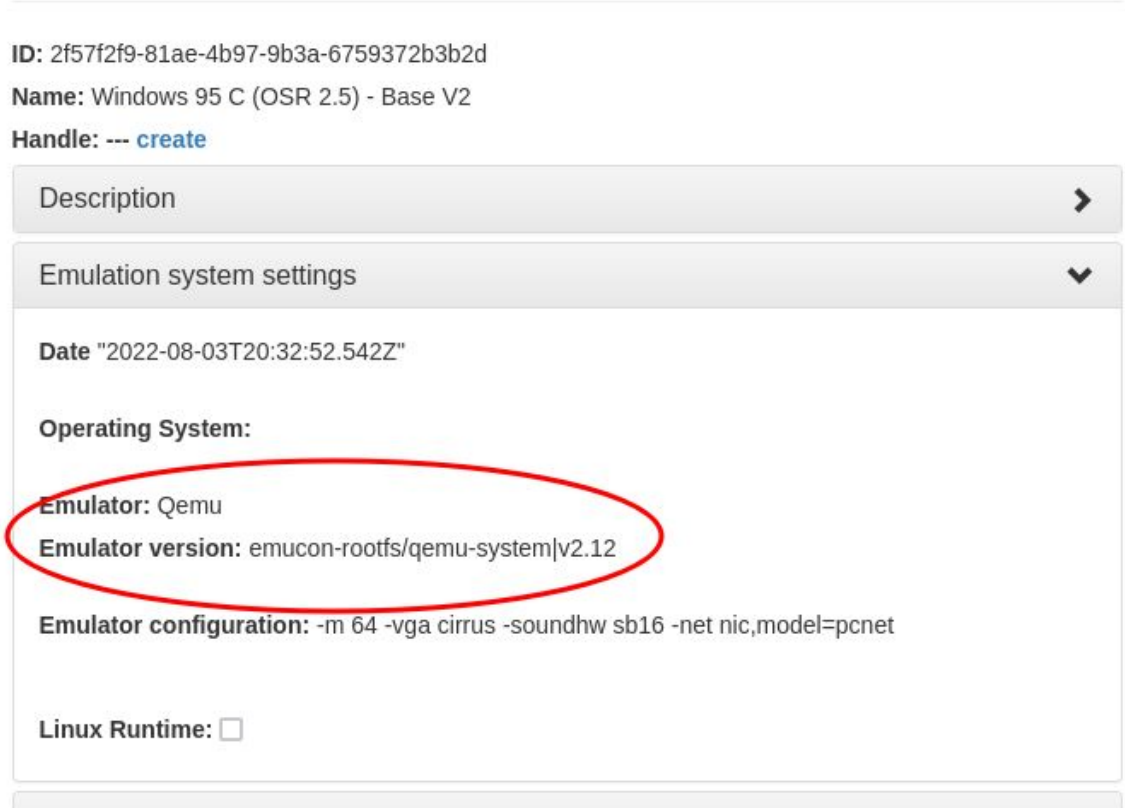
Step 4: Windows 95 was released in a particular *year*: ***1995***



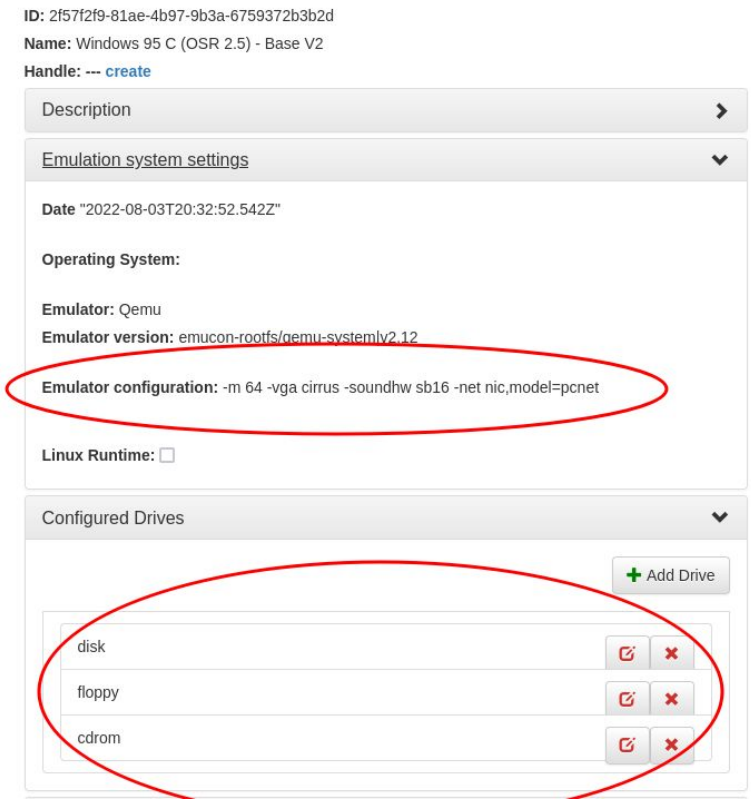

Step 5: Windows 95 was distributed using certain ***installation media***: ***CD-ROM***


Step 6: With a selected operating system version and year we can also assume the type of ***processor*** we need to emulate: ***32-bit x86*** *(*also known as ***i386****)*

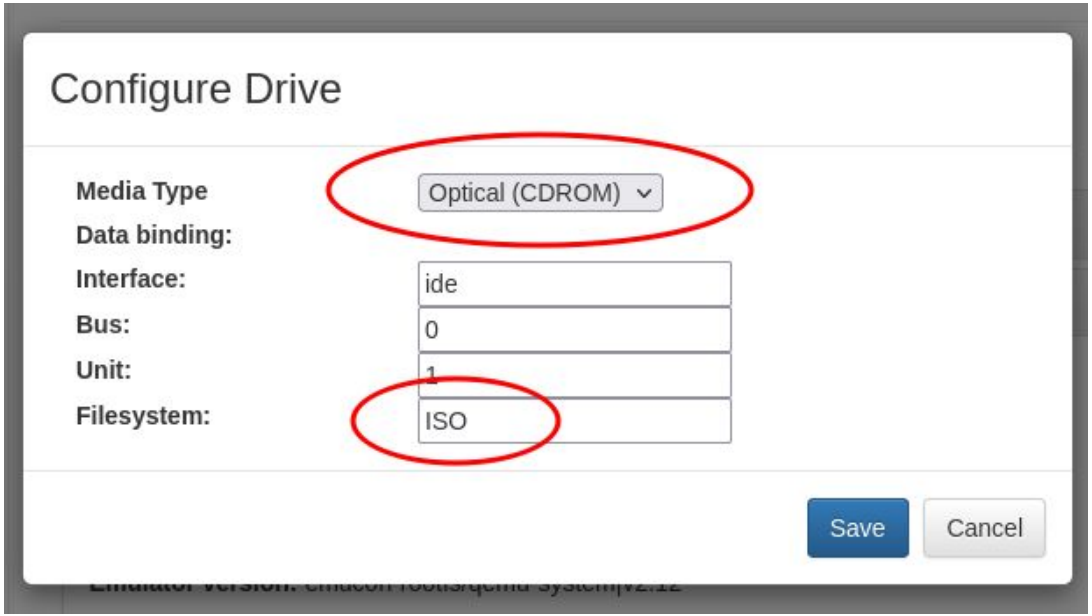

Step 7: This processor requires a particular ***emulator***: ***QEMU*** (specifically ***qemu-system-i386***)



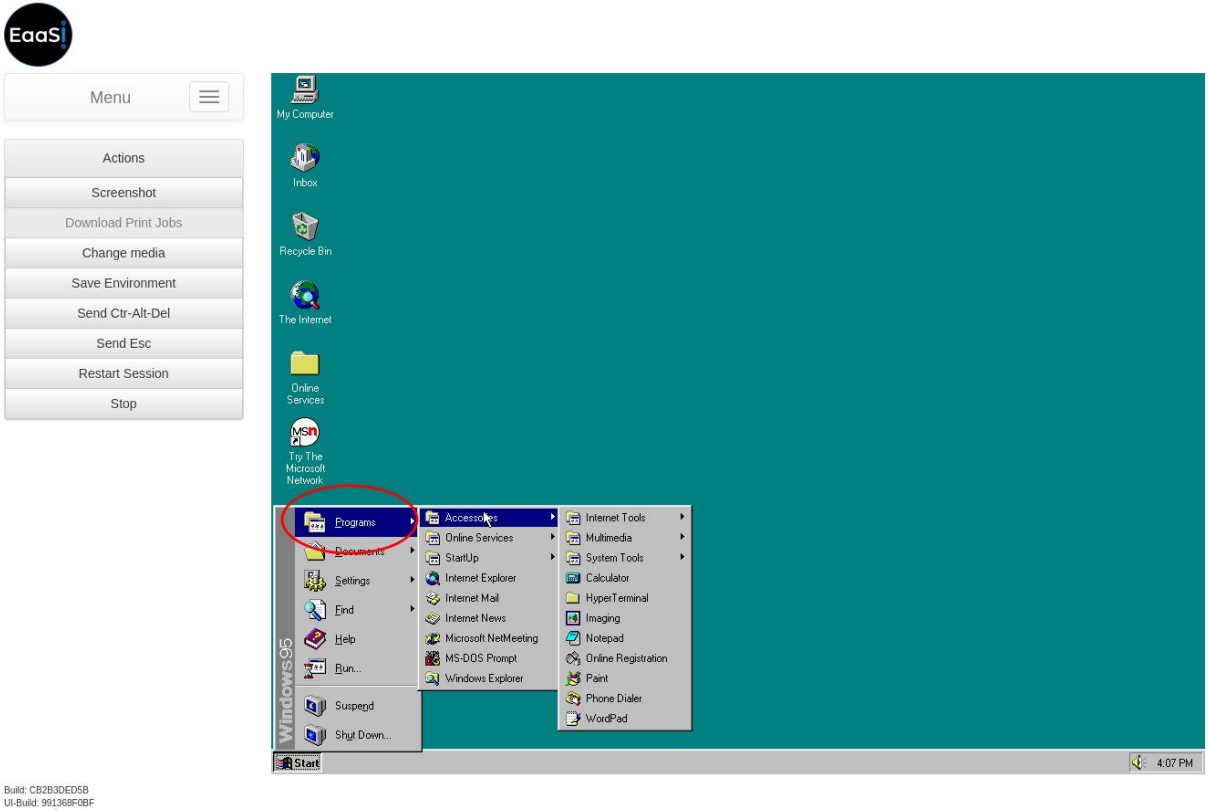

Step 8: By necessity we must also pick a particular ***version*** of the emulator as well: QEMU ***2.12***



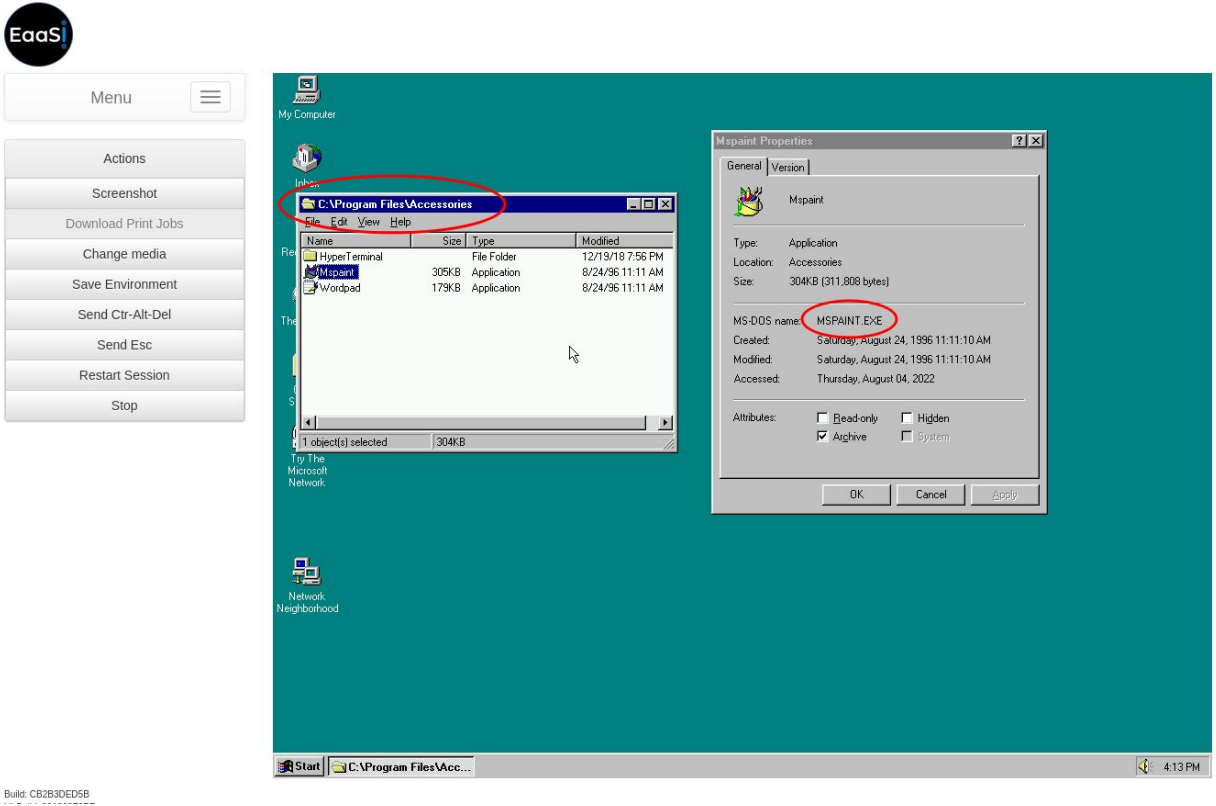

Step 9: This emulator version can recreate a set of ***hardware components***: a **generic PS/2 mouse and keyboard, 64 MB of RAM, Cirrus CLGD 5446 VGA card, Creative Soundblaster 16 sound card, AMD PCNET ethernet card, 2 PCI IDE interfaces** with **hard disk**, **floppy disk** and **CD-ROM drive** support



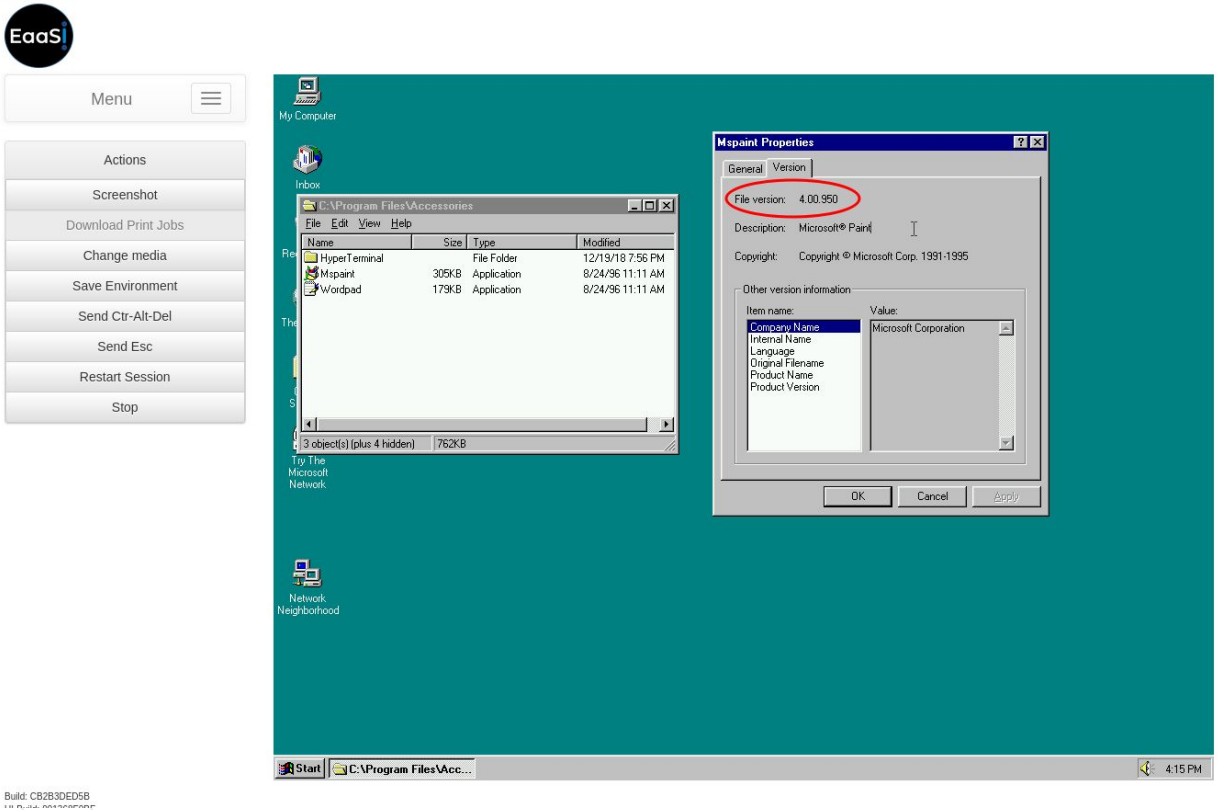

Step 10: QEMU's emulated CD-ROM drive can read certain ***types*** of ***installation media*** - like a disk image made from the original CD-ROM (such disk images are commonly referred to as ***ISOs***)



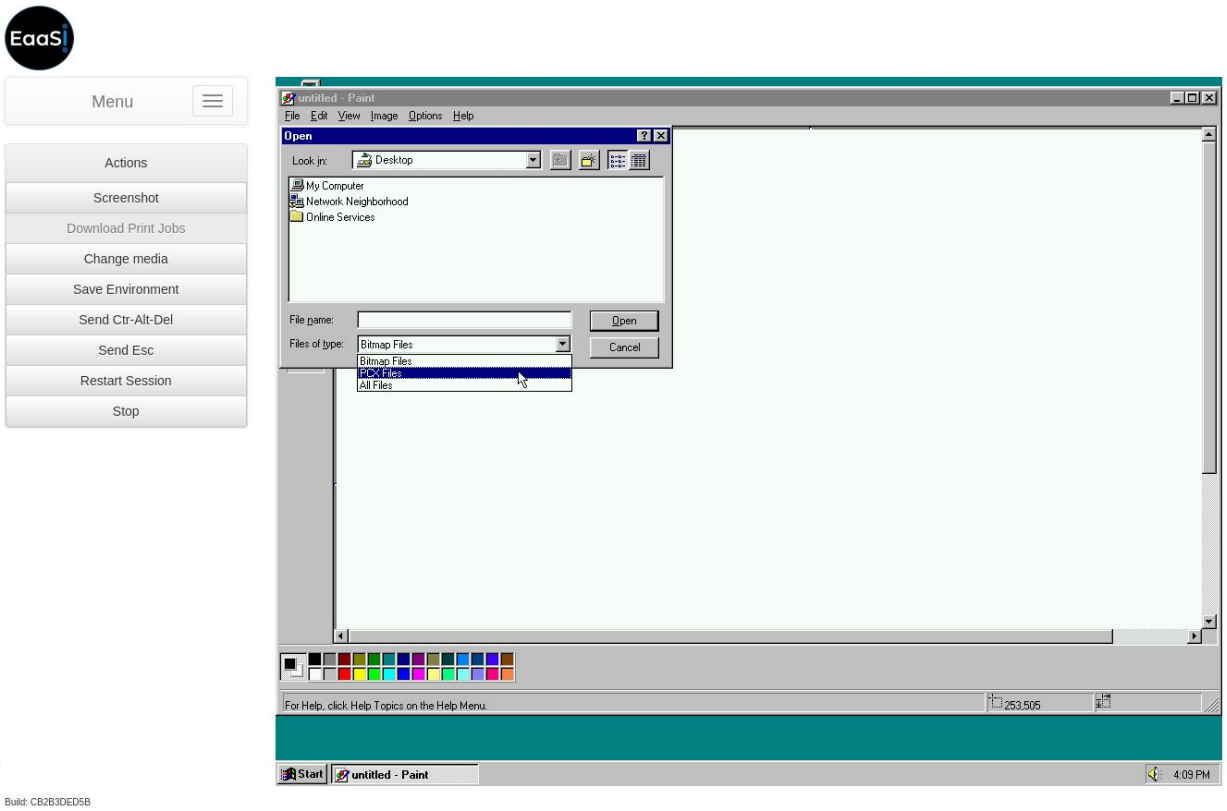

Step 11: The Windows 95 ISO installs a list of ***software products*** that make up the operating system - ***Wordpad, Paint, Calculator, File Manager, Internet Explorer***, etc.



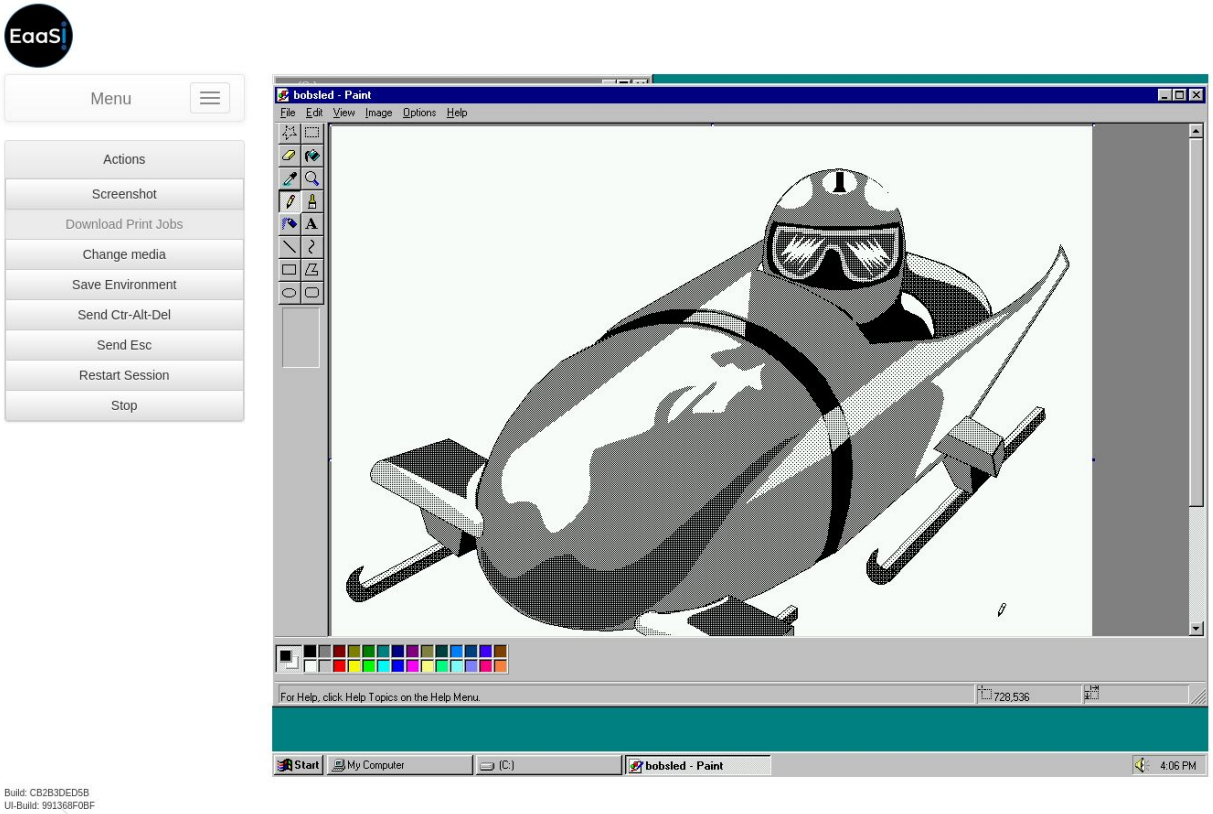

Step 12: Microsoft Paint is opened by running an ***executable*** at a certain ***location - C:\Program Files\Accessories\Mspaint.exe***




Step 13: Microsoft Paint has its own ***version*** - Microsoft Paint ***4.00.950***



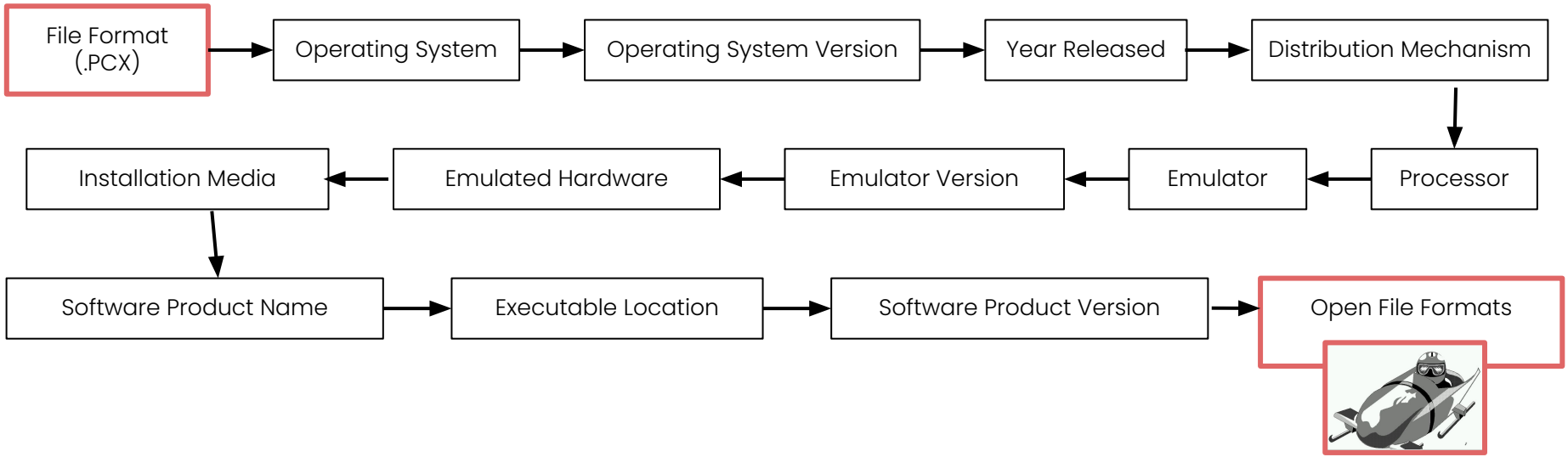

Step 14: This particular version of this particular software can ***open*** certain ***file formats*** - including ***.PCX*** files (like the one we started with!)




Step 15: Finally, thanks to our emulation of Windows 95 and Microsoft Paint 4.00.950, I can now open my original .PCX file and confirm - it's an ***image of a bobsled***!

# Thanks, Metadata!

- Various pieces of software metadata allowed us to emulate and render this file as accurately as possible
- Pushing standards and investigating automation will help us speed the journey



# Credits

- Training Module written and designed by Ethan Gates, Software Preservation Analyst, Yale University Library & Claire Fox, Metadata Analyst, Yale University Library
- Original photos, screenshots, and videos recorded by Ethan Gates
- Icons sourced from The Noun Project
- EaaSI program of work sponsored by the Alfred P. Sloan Foundation and the Mellon Foundation, hosted by Yale University Library

Yale

*Principle Partner*

ALFRED P. SLOAN FOUNDATION

*Sponsor*

Mellon Foundation

*Sponsor*