Harvard Library

2D and 3D Format Selection and Metadata Analysis

FINAL REPORT

Submitted By:

Drexel University

3141 Chestnut Street,
Philadelphia, PA 19104

June 30th, 2015

Isaac Simmons

Primary Objectives

The chief objective is to assist in identifying 2D and 3D file formats for acceptance into and preservation within the Harvard Digital Repository Service. We will accomplish this by:

- Surveying available formats in this domain and describing the best candidates for curation.
- Identifying useful metadata for extraction and mapping that metadata into larger schema.
- Integrating these results with the tools and policies surrounding ingest of a new dataset into the DRS.

Work will be conducted in collaboration with Harvard Libraries and weekly teleconferences will be held to report progress updates and receive feedback on deliverables. The overall project will be organized into 4 main phases:

1. Format Analysis
2. Metadata Analysis
3. Content Modeling
4. Tool Analysis

Project Status

High Level Summary

The allotted hours have been exhausted and work on the project is completed as of June 30th, 2015. Phase 1 and Phase 2 tasks are completed, but outstanding deliverables remain in the Phase 3 and 4 tasks. See below for status of specific deliverables.

Task Details

Phase 1: Format Analysis

Intermediate deliverables (candidate format list, format properties) have been rolled into the 2D and 3D format matrix and form the basis for the selected rows and columns. Additionally, the "acceptance summary" row summarizes the recommended selections. The "Format Profiles" are what was referred to as "Format Descriptions" in the list of deliverables.

All products are available online in Google drive, and selected documents are included in the appendices of this report.

Phase 2: Metadata Analysis

Metadata analysis has been concluded. The inclusion of the document MD fields is recommended, and for additional Computer Aided Drafting (CAD) specific fields, no suitable existing schema was found, so the "cad-md" schema is proposed. For most deliverables in this phase, the 2D and 3D products have been combined into a single document which then makes that distinction rather than including two separate schema.

Recommendations, schema documentation, the XSD schema, and example outputs associated with a few specific sample files are included both on Google Drive and in the appendices of this report.

Phase 3: Content Modeling

Content modeling efforts are ongoing. Preliminary products and a draft content model description are available in Google Drive.

Phase 4 Tool Analysis

Tool analysis efforts are ongoing. Sample files have been gathered and preliminary FITS identification results on those files have been evaluated. A survey of possible tools for integration with the DRS for either file identification, metadata extraction, or format conversion has been begun.

FITS has been tested against all of the sample files and those outputs are available in Google Drive.

Conferences Meetings and Demonstrations

Meeting	Location	Date
Weekly Status Meetings	Skype	Weekly, Variable
Kickoff Meeting	Cambridge, MA	January 26th, 2015

Issues

Some Task 3 and 4 Deliverables remain uncompleted. A separate proposal will be developed to cover these.

Deliverables

Name	Description	Phase	Status
Weekly Updates	Verbal updates given in teleconferences	ALL	Completed
2D Candidate List	Document	1	Column 1 of Format Matrices (below)
3D Candidate List	Document	1	
2D Format Properties	Document	1	Row 1 Labels in Format Matrices (below)
3D Format Properties	Document	1	
2D Format Matrix	Access to document in Google Drive	1	“2D Matrix” in Google Drive
3D Format Matrix	Access to document in Google Drive	1	“3D Matrix” in Google Drive
2D Class A and B Descriptions	Access to document in Google Drive	1	“Acceptance Summary” row in Format Matrices and Format Profiles (Google Drive, Appendix A)
3D Class A and B Descriptions	Access to document in Google Drive	1	
2D Metadata Elements	Document	2	Combined with Schema Recommendations
3D Metadata Elements	Document	2	

			(below)
2D Schema Recommendations	Document	2	“Metadata Recommendations” and “Metadata Recommendations – Schema” in Google Drive or Appendix B, C
3D Schema Recommendations	Document	2	
2D Sample Files and Metadata	Access to online file collection	2	“Sample Files” subdirectory and “Metadata Recommendations – Sample Files” in Google Drive or Appendix D
3D Sample Files and Metadata	Access to online file collection	2	
2D and 3D Content Models	Document	3	Incomplete.
FITS current performance report	Spreadsheet	4	Incomplete
FITS Configuration and Tool Recommendations	Document (access to software if required for recommendations)	4	Incomplete
FITS Schema Recommendations	Updated schema file plus descriptive document	4	Incomplete
Sample Files and FITS output	Access to online file collection	4	“Sample Files” and “sample_results” subdirectories in Google Drive

All documents have also been delivered via. Google Drive in the shared folder named “2D3DFormats”

Personnel and Contacts

Drexel Faculty

Jane Greenberg

Drexel Staff

Isaac Simmons
Research Engineer

Adrian Ogletree
Research Program Manager – Metadata Research Center

Colleen Kavanaugh
Program Administrator

Harvard

Andrea Goethals

Appendix A - Format Profiles

AutoCAD Drawing Format Profile - 2D/3D

Full name (taken from the specification if applicable) and common aliases

AutoCAD Drawing, AutoCAD Drawing Database File, DWG, AutoCAD .dwg File, AutoDesk's Drawing File

Brief description

The .dwg file format is one of the most commonly used design data formats, found in nearly every design environment. It signifies compatibility with AutoCAD technology. Autodesk created .dwg in 1982 with the launch of its first version of AutoCAD software. There have been 19 revisions of the format since then, the latest in 2013.

There are several claims to control of the DWG format. As the biggest and most influential creator of DWG files it is Autodesk who designs, defines, and iterates the DWG format as the native format for their CAD applications. Autodesk sells a read/write library, called RealDWG, under selective licensing terms for use in non-competitive applications. Several companies have attempted to reverse engineer Autodesk's DWG format, and offer software libraries to read and write Autodesk DWG files. The most successful is OpenDWG / Open Design Alliance, a non-profit consortium created in 1998 by a number of software developers, released a read/write/view library called the OpenDWG Toolkit.

It is the native format for several CAD packages including DraftSight, AutoCAD, IntelliCAD (and its variants), Caddie and Open Design Alliance compliant applications . In addition, DWG is supported non-natively by many other CAD applications.

Structure of the current DWG format

- Header
  - Version
  - Magic Number
  - Simple metadata

  Classes, Objects, Images
  - Encoded binary data
  - Checksums

Key adopters of the format (e.g. large repositories or academic libraries, domains)

DWG is among the most widely used CAD formats in the field. It is a preferred preservation format for CAD data (2D and 3D) with the Archaeology Data Service and is a preferred format for CAD data by the Library and Archives of Canada.

Applicable MIME media types

application/acad, application/x-acad, application/autocad_dwg, image/x-dwg, application/dwg, application/x-dwg, application/x-autocad, image/vnd.dwg, drawing/dwg

Applicable file extensions

.dwg

The organization/individual/company that originally developed it

Autodesk

The organization/individual/company that currently maintains it

Autodesk, Open Design Alliance

Availability and location of specifications (direct URLs if available)

Not officially available from Autodesk

Available from Open Design Alliance

http://www.opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf

Brief information about patent/license issues

Proprietary. Multiple claims to control. Autodesk licenses RealDWG libraries. However, the Open Design Alliance also maintains the OpenDWG toolkit.

Key related links (Websites describing it, documentation, etc.)

- http://fileinfo.com/extension/dwg
- http://en.wikipedia.org/wiki/.dwg
- http://www.opendesign.com/
- http://www.autodesk.com/products/dwg

Risk summary

- **Proprietary Format**
    - All generations are proprietary, with single vendor support, closed source, closed specification, and almost no viable open source implementations at this time

  Encryption
    - In AutoCAD 2004 version files and later, password protection can be enabled

  New Versions
    - New versions are published about once a year
    - While backwards compatibility remains good, software will need to be updated to deal with ingest of newer versions

Mitigation of key risks

- **Mitigating proprietary format risk**
    - Convert to other formats:
        - Convert DWG artifacts to other more open formats (though some detail may be lost in the process), keeping originals
    - Access via emulation:
        - Maintain copies of free viewing and conversion tools in an emulation environment

  Mitigating encryption risks
    - Disallow password protected files
    - Remove encryption and re-save files

  Mitigating new versions risk
    - Periodically update software tools for dealing with DWG files in archive
    - Re-save files in latest versions when possible, keeping originals
    - (Current) strong backwards compatibility for reading files means this is not an immediate danger

References

Archaeology Data Service. (n.d.). *Archaeology Data Service / Digital Antiquity Guides to Good Practice*. Retrieved from http://guides.archaeologydataservice.ac.uk/g2gp/Cad_3-2, http://guides.archaeologydataservice.ac.uk/g2gp/LaserScan_3-1

Library and Archives Canada. (n.d.). *Guidelines on File Formats for Transferring Information Resources of Enduring Value*. Retrieved from http://www.bac-lac.gc.ca/eng/services/government-information-

resources/guidelines/Pages/guidelines-file-formats-transferring-information-resources-enduring-value.aspx#u

Drawing Interchange Format Profile - 2D/3D

Full name (taken from the specification if applicable) and common aliases

Drawing eXchange Format, Drawing Interchange Format, AutoCAD DXF (AutoCAD DXB, Drawing eXchange Binary)

Brief description

AutoCAD DXF is a CAD data file format developed by Autodesk for enabling data interoperability between AutoCAD and other programs.

DXF was originally introduced in December 1982 as part of AutoCAD 1.0, and was intended to provide an exact representation of the data in the AutoCAD native file format, DWG, for which Autodesk for many years did not publish specifications. Because of this, correct imports of DXF files have been difficult. Autodesk now publishes the DXF specifications as a PDF on its website.

Versions of AutoCAD from Release 10 (October 1988) and up support both ASCII and binary forms of DXF. Earlier versions support only ASCII. DXB is the binary version of a DXF file, which is text-based. DXB files are smaller and load faster than DXF files, but are not as compatible with other programs as DXF files are.

As AutoCAD has become more powerful, supporting more complex object types, DXF has become less useful. Certain object types, including ACIS solids and regions, are not documented. Other object types, including AutoCAD 2006's dynamic blocks, and all of the objects specific to the vertical market versions of AutoCAD, are partially documented, but not well enough to allow other developers to support them. For these reasons many CAD applications use the DWG format which can be licensed from AutoDesk or non-natively from the Open Design Alliance.

Structure of the current DXF format

- Header
- Classes
- Tables
- Blocks
- Entities
- Objects
- Thumbnail Image

Key adopters of the format (e.g. large repositories or academic libraries, domains)

DXF is a preferred preservation format for CAD data (2D and 3D) with the Archaeology Data Service and is a preferred format for CAD data by the Library and Archives of Canada.

Applicable MIME media types

application/dxf, application/x-autocad, application/x-dxf, drawing/x-dxf, image/vnd.dxf, image/x-autocad, image/x-dxf, zz-application/zz-winassoc-dxf, (application/dxb, application/x-dxb, drawing/x-dxb, image/x-dxb)

Applicable file extensions

.dxf, (.dxb)

The organization/individual/company that originally developed it

Autodesk

The organization/individual/company that currently maintains it

Autodesk

Availability and location of specifications (direct URLs if available)

Available directly from from Autodesk

http://images.autodesk.com/adsk/files/acad_dxf0.pdf

Brief information about patent/license issues

Proprietary. Controlled by Autodesk but freely published online.

Key related links (Websites describing it, documentation, etc.)

- http://fileinfo.com/extension/dxf
- http://en.wikipedia.org/wiki/AutoCAD_DXF
- http://www.autodesk.com/techpubs/autocad/acad2000/dxf/dxf_format.htm

Risk summary

- **Proprietary Format**
  - All generations are proprietary, with single vendor support
- **New Versions**
  - New versions are published about once a year
  - While backwards compatibility remains good, software will need to be updated to deal with ingest of newer versions

Mitigation of key risks

- **Mitigating proprietary format risk**
  - Convert to other formats:
    - Convert DXF artifacts to other more open formats (though some detail may be lost in the process), keeping originals
  - Access via emulation:
    - Maintain copies of free viewing and conversion tools in an emulation environment

  Mitigating new versions risk
  - Periodically update software tools for dealing with DWG files in archive
  - Re-save files in latest versions when possible, keeping originals
  - (Current) strong backwards compatibility for reading files means this is not an immediate danger

References

Archaeology Data Service. (n.d.). *Archaeology Data Service / Digital Antiquity Guides to Good Practice*. Retrieved from http://guides.archaeologydataservice.ac.uk/g2gp/Cad_3-2, http://guides.archaeologydataservice.ac.uk/g2gp/LaserScan_3-1

Library and Archives Canada. (n.d.). *Guidelines on File Formats for Transferring Information Resources of Enduring Value*. Retrieved from http://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Pages/guidelines-file-formats-transferring-information-resources-enduring-value.aspx#u

PDF/A Format Profile - 2D

Full name (taken from the specification if applicable) and common aliases

PDF/A, PDF/A-1, PDF/A-2, PDF/A-3, Use of PDF 1.4 (PDF/A-1), Use of ISO 32000-1 (PDF/A-2), Use of ISO 32000-1 with support for embedded files (PDF/A-3)

Brief description

PDF/A is a subset of PDF that eliminates certain risks threatening the one-to-one future reproducibility of the content. PDF/A forbids dynamic content to ensure that the user sees the exact same content both today and for years to come. Everything that is required to render the document the exact same way, every time, is contained in the PDF/A file: fonts, colour profiles, images etc. PDF/A is also an ISO standard, guaranteeing that future software generations will know how to open and render PDF/A files.

Structure of the PDF/A-1 format

- Container: PDF 1.4
- Content:
    - PDF/A allows only a subset of the possible PDF elements
    - No external content references or fonts (all data must be embedded within the document itself)
    - No Javascript
    - No Audio/Video
    - Certain compression and image options are forbidden due to legal concerns
    - Encryption is forbidden

Structure of the PDF/A-2 format

- Similar to PDF/A-1 though a few additional content items are allowed
- Container is PDF 1.7 (itself an ISO Standard) instead of PDF 1.4

Structure of the PDF/A-3 format

- Similar to PDF/A-2
- Includes support for embedding additional files of arbitrary formats

Key adopters of the format (e.g. large repositories or academic libraries, domains)

PDF/A-1 is a NARA (National Archives and Records Administration) preferred format for scanned text, posters, presentations, and text.

PDF/A-2 is a NARA acceptable format for scanned text and presentations and a NARA preferred format for text.

PDF/A is a preferred format for textual works by the Library of Congress.

It is an archival format for texts and documents for the Archaeology Data Service.

It is a preferred format (PDF/A-1, PDF/A-2) of the Library and Archives of Canada for text, scanned text, and presentations.

PDF/A is an Archivematica preservation format for documents.

Applicable MIME media types

application/pdf

Applicable file extensions

.pdf

The organization/individual/company that originally developed it

Adobe

The organization/individual/company that currently maintains it

Adobe

Availability and location of specifications (direct URLs if available)

ISO 19005-1/19005-2/19005-3
http://www.iso.org/iso/catalogue_detail?csnumber=38920

http://www.iso.org/iso/catalogue_detail.htm?csnumber=50655

http://www.iso.org/iso/catalogue_detail.htm?csnumber=57229

Brief information about patent/license issues

ISO Standard, Controlled by Adobe

Key related links (Websites describing it, documentation, etc.)

- http://www.pdfa.org/2011/06/pdfa-faq/
- http://en.wikipedia.org/?title=PDF/A
- http://www.iso.org/iso/catalogue_detail?csnumber=38920
- http://www.iso.org/iso/catalogue_detail.htm?csnumber=50655
- http://www.iso.org/iso/catalogue_detail.htm?csnumber=57229

Risk summary

- **Limited Capabilities**
  - PDF/A files cannot contain embedded 3D objects

  Multiple Profiles
  - PDF/A-1, PDF/A-2, and PDF/A-3 all differ in their capabilities and restrictions
  - Furthermore, there are PDF/A-1a and PDF/A-1b, PDF/A-2a, PDF/A-2b, and PDF/A-2u compliance levels

Mitigation of key risks

- **Mitigating limited capabilities**
  - For files that include 3D data, consider using PDF/E or another format instead

  Mitigating multiple profiles
  - Prefer only PDF/A-1 or possibly PDF/A-2

References

Archaeology Data Service. (n.d.). *Archaeology Data Service / Digital Antiquity Guides to Good Practice*. Retrieved from http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_2

Archivematica. (2014). *Format Policies*. Retrieved from https://www.archivematica.org/wiki/Format_policies

Library of Congress. (n.d.). *Recommended Format Specifications. Textual Works and Musical Compositions*. Retrieved from http://www.loc.gov/preservation/resources/rfs/textmus.html

NARA. (n.d.). *NARA 2014-04: Appendix A, Revised Format Guidance for the Transfer of Permanent Electronic Records – Tables of File Formats*. Retrieved from http://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

PDF/E Format Profile - 2D/3D

Full name (taken from the specification if applicable) and common aliases

Portable Document Format, PDF/E Standard, PDF/E-1

Brief description

PDF/E (ISO 24517-1:2008), a document standard ratified by ISO in 2007, evolved from the need for an open, neutral exchange format for engineering and technical documentation. While multiple proprietary formats exist, they each have their own viewers, making it difficult to repurpose 3D and engineering data for downstream uses. The cost of distributing and storing paper contributes to the high cost of managing distribution and change throughout the project for product development teams as well as for extended supply chains. Like PDF, PDF/E is a digital container which supports a wide variety of content and can be viewed and marked up using free and widely available Adobe Reader® software. PDF/E can help support the secure distribution of sensitive information and reduce the complexity and costs associated with distributing and storing paper. While PDF/E is an open standard developed and maintained by an ISO working group, it also leverages U3D, another open standard, for the representation of 3D content.

ISO 24517-1:2008 specifies the use of the Portable Document Format (PDF) Version 1.6 for the creation of documents used in engineering workflows. It provides specifications for the creation, viewing, and printing of documents used in engineering workflows. PDF/E facilitates the exchange of documentation and drawings to share with others in the supply chain or streamline review and markup. It specifies PDF settings suitable for building, manufacturing, and geospatial workflows and supports interactive media, including animation and 3D.

Structure of the PDF/E format

- Container: PDF
  - Based on PDF 1.6

  Content:
  - U3D Content
  - Text
  - Many embedded content types allowed

Key adopters of the format (e.g. large repositories or academic libraries, domains)

PDF is a very popular format, though the PDF/E specification is fairly new. It is included in NARA's list of "Acceptable" formats for CAD data.

Applicable MIME media types

application/pdf

Applicable file extensions

.pdf

The organization/individual/company that originally developed it

Adobe

The organization/individual/company that currently maintains it

Adobe

Availability and location of specifications (direct URLs if available)

ISO 24517-1

http://www.iso.org/iso/catalogue_detail.htm?csnumber=42274

Brief information about patent/license issues

ISO Standard

Key related links (Websites describing it, documentation, etc.)

- http://en.wikipedia.org/wiki/PDF/E
- http://www.iso.org/iso/catalogue_detail.htm?csnumber=42274
- http://www.aiim.org/documents/standards/PDF-E/PDF_E_FAQ-Edits_Jan.pdf

Risk summary

- **Not developed as an archiving format**
    - This format was not specifically designed as an archival format
    - However PDF/E and U3D are both open formats and may be suitable

  New Versions
    - PDF/E-2 is currently under development
    - PDF/E-2 is expected to encourage PRC embedded 3D objects over U3D ones (though will support both)
- **Encryption**
    - PDF/E files may be encrypted

  Embedded Resources
    - Resources can be embedded within the main object, each of which may present new preservation issues.

Mitigation of key risks

- **Mitigating archival format risk**
    - Encourage creators to apply (some) principles of PDF/A-1 to PDF/E documents

  Mitigating new version risk
    - PDF/E revisions are being developed in such a way that existing PDF/E-1 documents will be valid with future versions of the specification
    - Access via emulation:
        - Providing legitimate copies of OS X and Pages can be found.
        - Further complicated by the fact that current versions of Pages are distributed via the AppStore and so older versions are not easily available.

  Mitigating encryption risk
    - Examine incoming material to check for the presence of encryption, reject submission or decrypt prior to storage

  Mitigating embedded resources risk
    - Limit use of embedded multimedia components in document

References

Archaeology Data Service. (n.d.). *Archaeology Data Service / Digital Antiquity Guides to Good Practice*. Retrieved from http://guides.archaeologydataservice.ac.uk/g2gp/VectorImg_2

Archivematica. (2014). *Format Policies*. Retrieved from https://www.archivematica.org/wiki/Format_policies

Library and Archives Canada. (n.d.). *Guidelines on File Formats for Transferring Information Resources of Enduring Value*. Retrieved from http://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Pages/guidelines-file-formats-transferring-information-resources-enduring-value.aspx#u

Library of Congress. (n.d.). *Recommended Format Specifications. Textual Works and Musical Compositions*. Retrieved from http://www.loc.gov/preservation/resources/rfs/textmus.html

NARA. (n.d.). *NARA 2014-04: Appendix A, Revised Format Guidance for the Transfer of Permanent Electronic Records – Tables of File Formats*. Retrieved from http://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#computeraided

Wikipedia. (n.d.). *Pages (word processor) - Version history*. Retrieved from http://en.wikipedia.org/wiki/Pages_(word_processor)_-_Version_history

STEP-File Profile - 2D/3D

Full name (taken from the specification if applicable) and common aliases

STEP-File, p21 File, STEP Physical File

(STEP = ISO 10303 = Standard for the Exchange of Product Model Data = Automation systems and integration — Product data representation and exchange)

Brief description

STEP-File is the most widely used data exchange form of STEP. ISO 10303 can represent 3D objects in Computer-aided design (CAD) and related information. Due to its ASCII structure it is easy to read with typically one instance per line. The format of a STEP-File is defined in ISO 10303-21 *Clear Text Encoding of the Exchange Structure*.

ISO 10303-21 defines the encoding mechanism on how to represent data according to a given EXPRESS schema, but not the EXPRESS schema itself. A STEP-File is also called p21-File and STEP Physical File. The file extensions .stp and .step indicates that the file contain data conforming to STEP Application Protocols while the extension .p21 should be used for all other purposes.

Structure of the STEP-File format

STEP-File is an ASCII format

Content:

- Header
  - “Magic Number”
  - File metadata/Description
  - Schema used in data section
- Data
  - Data objects
  - Mappings between data objects

Key adopters of the format (e.g. large repositories or academic libraries, domains)

STEP is a NARA Preferred preservation format for CAD data and an acceptable preservation format for the Library and Archives of Canada.

Applicable MIME media types

application/step

Applicable file extensions

.step, .stp

The organization/individual/company that originally developed it

ISO

The organization/individual/company that currently maintains it

ISO

Availability and location of specifications (direct URLs if available)

ISO Standard

http://www.iso.org/iso/home/catalogue_detail.htm?csnumber=33713

Brief information about patent/license issues

Open Standard.

Key related links (Websites describing it, documentation, etc.)

- http://en.wikipedia.org/wiki/ISO_10303-21
- http://en.wikipedia.org/wiki/ISO_10303
- http://www.iso.org/iso/home/catalogue_detail.htm?csnumber=33713

Risk summary

- **Vendor Adoption**
  - Support for full set of features varies between software tools
  - Most often, this is an import/export format, not the native format of the software

Mitigation of key risks

- **Mitigating vendor adoption**
  - These concerns apply more to people producing the artifacts than the preservation of those files
  - Ensure availability of open source tools for curation

References

Library and Archives Canada. (n.d.). *Guidelines on File Formats for Transferring Information Resources of Enduring Value*. Retrieved from http://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Pages/guidelines-file-formats-transferring-information-resources-enduring-value.aspx#u

NARA. (n.d.). *NARA 2014-04: Appendix A, Revised Format Guidance for the Transfer of Permanent Electronic Records – Tables of File Formats*. Retrieved from http://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#computeraided

Wikipedia (n.d.) *ISO 10303-21.* Retrieved from http://en.wikipedia.org/wiki/ISO_10303-21

X3D Format Profile - 3D

Full name (taken from the specification if applicable) and common aliases

Information technology — Computer graphics and image processing — Extensible 3D (X3D), Extensible 3D Graphics

Brief description

X3D is a royalty-free open standards file format and run-time architecture to represent and communicate 3D scenes and objects using XML. It is an ISO ratified standard () that provides a system for the storage, retrieval and playback of real time graphics content embedded in applications, all within an open architecture to support a wide array of domains and user scenarios.

X3D has a rich set of componentized features that can tailored for use in engineering and scientific visualization, CAD and architecture, medical visualization, training and simulation, multimedia, entertainment, education, and more.

The development of real-time communication of 3D data across all applications and network applications has evolved from its beginnings as the Virtual Reality Modeling Language (VRML) to the considerably more mature and refined X3D standard.

X3D strives to become the 3D standard for the Web, as integrated in the HTML5 pages as other XML dialects (MathML, SVG) already are there.

Structure of the X3D format

- Container: XML
- Content:
  - A wide variety of content can be embedded, including Javascript, 2D and 3D objects, textures, etc
  - Hyperlinked data can be included that must be loaded from external locations

Key adopters of the format (e.g. large repositories or academic libraries, domains)

X3D is a NARA (National Archives and Records Administration) preferred format for CAD data.

Applicable MIME media types

model/x3d+xml, model/x3d+binary, model/x3d+vrml

Applicable file extensions

.x3d, .x3dv, .x3db, .x3dz, .x3dbz, .x3dvz (vrml, binary, compressed/zipped)

The organization/individual/company that originally developed it

ISO

The organization/individual/company that currently maintains it

ISO

Availability and location of specifications (direct URLs if available)

ISO 19775/19776/19777

http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=60760

http://www.web3d.org/documents/specifications/19775-1/V3.2/index.html

Brief information about patent/license issues

Royalty-free ISO standard

Key related links (Websites describing it, documentation, etc.)

- http://www.web3d.org/x3d/what-x3d
- http://en.wikipedia.org/wiki/X3D
- http://fileformats.archiveteam.org/wiki/X3D

Risk summary

- **External References**
  - Externally referenced content may be difficult to identify, collect and preserve
  - X3D documents can embed network-accessible content within a file

  Multiple Profiles
  - X3D specifies several profiles for varying levels of capability including X3D Core, X3D Interchange, X3D Interactive, X3D CADInterchange, X3D Immersive, and X3D Full
  - X3D supports multiple encodings (XML, VRML, Binary)
  - X3D supports optional compression
  - Support for all of these options may vary across software platforms

  Embedded
  - X3D documents may themselves be embedded in other XML or HTML files which present their own archival challenges

Mitigation of key risks

- **Mitigating external references risk**
  - Disallow X3D files that include linked external content
  - Download linked external content and embed directly in X3D file

  Mitigating multiple profiles risk
  - Weigh different profiles and decide if only some will be accepted
  - Convert to a preferred encoding

  Mitigating embedded
  - Disallow X3D files embedded in other documents
  - Make sure that containing documents themselves conform to archival requirements

References

Archaeology Data Service. (n.d.). *Archaeology Data Service / Digital Antiquity Guides to Good Practice*. Retrieved from http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_2

Archivematica. (2014). *Format Policies*. Retrieved from https://www.archivematica.org/wiki/Format_policies

Library of Congress. (n.d.). *Recommended Format Specifications. Textual Works and Musical Compositions*. Retrieved from http://www.loc.gov/preservation/resources/rfs/textmus.html

NARA. (n.d.). *NARA 2014-04: Appendix A, Revised Format Guidance for the Transfer of Permanent Electronic Records – Tables of File Formats*. Retrieved from http://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#computeraided

Appendix B - Metadata Recommendations

Table of Contents

Recommendations

Goals:

- Evaluate the completeness of the content after transformations (e.g. number of geometric primitives).
- Aid in selecting or aggregating files for risk analysis, preservation or delivery planning

Certain elements are applicable to word documents that may also be applicable to CAD and similar file types. DocMD (document metadata) includes PageCount, Language, Font, FontName, IsEmbedded, Reference, Features, documentMetadataExtension. For more formal definitions of these properties, see Chou & Goethals, 2012.

The DocMD elements that seem relevant to 2D and 3D files are:

Semantic Unit	PageCount
Semantic Components	None
Description	Total number of pages in the CAD file
Data Constraint	Min 1
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	

Semantic Unit	Language

Semantic Components	None
Description	A language identifier specifying the natural language used in the document
Data Constraint	String (or some kind of controlled vocabulary like ISO 639-2 alpha-3 language codes)
Obligation	Optional
Cardinality	0 - N
Characteristic	Content
Note	

Semantic Unit	Font
Semantic Components	FontName isEmbedded
Description	A list of fonts used in the document
Data Constraint	Container
Obligation	Optional
Cardinality	0 - N
Characteristic	Content, Appearance
Note	This element allows a repository to store the names of all fonts used in a file. Some repositories may choose to store only the non-embedded fonts. The use of non-embedded fonts may hinder the long term preservation of the documents. For example, a document encoded with a proprietary non-embedded math font may not be migrated due to unavailability of the specific math font. It is recommended that repositories record at least the nonembedded fonts to assist in identifying the documents with potential long-term preservation risks.

Semantic Unit	FontName
Semantic Components	None
Description	Name of a font
Data Constraint	String
Obligation	Mandatory
Cardinality	1

Characteristic	Content, Appearance
Note	

Semantic Unit	isEmbedded
Semantic Components	None
Description	An indication of whether or not a font is embedded in a document
Data Constraint	Boolean
Obligation	Optional
Cardinality	1
Characteristic	Content, Appearance
Note	

The following list includes general suggestions for the level of detail that should be translated between different CAD file formats (Wikipedia, CAD). Some of these are already covered by DocMD elements above.

- model description
- is the data wireframe, surface, or solid?
- topology (BREP) information
- face and edge identifications
- feature information and history
- PMI annotation
- text and annotations (fonts, format)
- color and layer of graphical objects

Based on these recommendations and the info available in the headers, I would recommend the following elements beyond what is in DocMD.

Semantic Unit	Features
Semantic Components	None
Description	Additional document features
Data Constraint	hasAnnotations, hasKinematics, hasMaterialProperties, hasTolerances, hasTransparencies
Obligation	Optional
Cardinality	0 - N

Characteristic	Content
Note	

Semantic Unit	Representation
Semantic Components	RepresentationType ObjectCount
Description	Types of geometric primitives present in the model
Data Constraint	Container
Obligation	Optional
Cardinality	0 - N
Characteristic	Content
Note	

Semantic Unit	RepresentationType
Semantic Components	
Description	Which method the file uses for representing shapes.
Data Constraint	has2DPointSets has3DPointSets has2DRasterData: Object contains two dimensional raster data or textured surfaces in a three dimensional model has3DRasterData: Object contains voxel data or stacked 2d raster data hasBREP: Object contains shapes defined by boundary representation methods hasImplicitCurves: Object contains two dimensional lines or curves defined by implicit equations hasImplicitSurfaces: Object contains three dimensional surfaces defined by implicit equations hasParametricCurves: Object contains two dimensional lines or curves defined by parametric equations hasParametricSurfaces: Object contains three dimensional surfaces defined by parametric equations hasTriangleMesh
Obligation	Required
Cardinality	1
Characteristic	Content

Note	The possible values for this can inform whether or not the file is 2D or 3D. The representation types hasImplicitCurves, hasImplicitSurfaces, hasParametricCurves, hasParametricSurfaces may be too low level to be meaningful. In particular, it is unlikely that they will be used in conjunction with the accompanying ObjectCount variable.

Semantic Unit	ObjectCount
Semantic Components	None
Description	Number of geometric primitives of the given representation type (facets of a polygon mesh, points in a point cloud, etc)
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Content
Note	Can be useful for validating conversions.

Semantic Unit	Units
Semantic Components	None
Description	The type of unit system defined in the file.
Data Constraint	hasStandard hasMetric
Obligation	Optional
Cardinality	0 - N
Characteristic	Content
Note	Some people will use both unit systems in one file.

Semantic Unit	Extent
Semantic Components	Dimension
Description	An approximate maximum extent of all aggregated objects contained in the file

Data Constraint	None
Obligation	Optional
Cardinality	0-1
Characteristic	Content
Note	

Semantic Unit	Dimension
Semantic Components	axis magnitude units
Description	A one-dimensional component of the approximate maximum extent of all aggregated objects contained in the file
Data Constraint	axis: x,y,z magnitude: decimal (positive) units: string (km, inches, miles, m, etc -- some standard set to use here?)
Obligation	Optional
Cardinality	1-3
Characteristic	Content
Note	

Note: The FITS documentation states that the following technical non-domain-specific metadata are captured. Therefore, they are not part of my recommendations (FITS XML).

- copyrightBasis element
- copyrightNote element
- created element (file creation date)
- creatingApplicationName element (name of the software used to create the file)
- creatingApplicationVersion element (version of the software used to create the file)
- creatingos element (Operating system used to create the file)
- filepath element (full filepath to the file)
- filename element (name of the file)
- fslastmodified element (last modified date based on file system metadata)
- inhibitorType element (type of file inhibitor)
- inhibitorTarget element (what is being inhibited)
- lastmodified element (last modified date based on metadata embedded in the file)
- md5checksum element (MD5 value for the file)
- rightsBasis element
- size element (size of the file in bytes)

Validity

Typical validation properties for solid models include the volume, centre of gravity and calculated weight of each solid in the model. For solid and surface models, surface areas can be used. Another versatile technique is the use of a point cloud: this is where a large set of co-ordinates is calculated such that each co-ordinate lies on a surface in the model. The distribution of these points should not be random: they can be sparse across flat surfaces, but need to be denser where surfaces curve more steeply, and particularly dense along edges and corners (Ball, 2013).

Notes on Specific File Types

DWG

http://www.opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf

DXF

http://images.autodesk.com/adsk/files/acad_dxf0.pdf

STEPFile

Types, entities, rules and functions (http://www.steptools.com/support/stdev_docs/stpcad/html/index.html)

As each CAD system has its own method of describing geometry, both mathematically and structurally, there is always some loss of information when translating data from one CAD data format to another. The intermediate file formats are also limited in what they can describe, and they can be interpreted differently by both the sending and receiving systems.
It is therefore important when transferring data between systems to identify what needs to be translated.

If only the 3D model is required for the downstream process, then only the model description needs to be transferred. However, there are levels of detail. For example: is the data wireframe, surface, or solid; is the topology (BREP) information required; must the face and edge identifications be preserved on subsequent modification; must the feature information and history be preserved between systems; and is PMI annotation to be transferred.

With product models, retaining the assembly structure may be required.

If drawings need to be translated, the wireframe geometry is normally not an issue; however text, dimensions and other annotation can be an issue, particularly fonts and formats. No matter what data is to be translated, there is also a need to preserve attributes (such as color and layer of graphical objects) and text information stored within the files. Sometimes, however, there is a problem caused by too much information being preserved. An example are the constraints placed on designers arising out of the design intent-history captured in parametric design systems. The receiving system must provide designers with the design freedom to modify geometry without having to understand the history of, or undo, the design tree. (Wikipedia, 2015, CAD)

Glossary

BREP (Boundary Representation): A method of solid modelling where the solids are defined in terms of their boundaries (surfaces).

CAD (Computer-Aided Design)

Exchange format: A format that has been designed to be read and written by several different software applications with a minimum of loss. Exchange formats can be vendor neutral (as with IGES and STEP AP 203) or tied to a popular software product, though in the latter case they are typically different from the software's native format. AutoDesk, for example, maintains an exchange format called DXF, which is related to but distinct from DWG, the native file format of its AutoCAD product.

Feature: A feature in the modelling sense is a generic characteristic or shape with a certain significance, with implications for its relationship with other features and various other parametric constraints. Examples might include a curved blend between two surfaces (which will affect how the boundary behaves under stress) or a keyway (which will need to accommodate a matching key).

PMI (Product and Manufacturing Information): In the widest sense, this refers to the additional information needed to manufacture a part from the shape data present in a 2D drawing or 3D CAD model. At a minimum, it includes geometric dimensions and tolerances (which see) but may include other annotations, and specifications of finishes and materials.

Shape data: The points, lines, surfaces and solid objects making up the geometric information in a CAD model, but not the product and manufacturing information, parametric relationships/properties, feature semantics or construction history.

References

Adobe. (2008). PRC Format Specification. Retrieved from http://help.adobe.com/livedocs/acrobat_sdk/9/Acrobat9_HTMLHelp/API_References/PRCReference/PRC_Format_Specification/index.html

AIG. Engineering File Format Registry. http://gicl.cs.drexel.edu/index.php/Category:Engineering_format

Ashenfelder, Mike. (2014). Untangling the Knot of CAD Preservation. “The Signal.” The Library of Congress. Retrieved from http://blogs.loc.gov/digitalpreservation/2014/08/untangling-the-knot-of-cad-preservation/

Autodesk. (2007). DXF Reference. Retrieved from http://images.autodesk.com/adsk/files/acad_dxf0.pdf

Ball, Alex. (2013). Preserving Computer-Aided Design (CAD). Digital Preservation Coalition Technology Watch Report.
Retrieved from http://dx.doi.org/10.7207/twr13-02

Chou, C. C., & Goethals, A. (2012). Document Metadata: document technical metadata for digital preservation. Retrieved from http://library.harvard.edu/sites/default/files/documentMD_2012.pdf

Ecma International. (2007). Universal 3D File Format. Retrieved from http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-363%204th%20Edition.pdf

FACADE. (2009). Final Report for the MIT FACADE Project: October 2006 – August 2009. Retrieved from http://www.cvaa.be/sites/default/files/projecten/bijlagen/bib_3896_facade_final.pdf

FITS XML. Retrieved from http://projects.iq.harvard.edu/fits/fits-xml

ISO 10303-21:2002 Industrial automation systems and integration -- Product data representation and exchange -- Part 21: Implementation methods: Clear text encoding of the exchange structure.

Open Design Alliance. (2013). Open Design Specification for .dwg files, Version 5.3. Retrieved from http://www.opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf

PREMIS Editorial Committee. (2008). PREMIS Data Dictionary for Preservation Metadata. Retrieved from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

Web3D Consortium. (2015). Recommended Standards. Retrieved from http://www.web3d.org/standards

Wikipedia. (2015). CAD data exchange. Retrieved from http://en.wikipedia.org/wiki/CAD_data_exchange

Wikipedia. (2015). ISO 10303-21. Retrieved from http://en.wikipedia.org/wiki/ISO_10303-21

Appendix

These are the metadata fields I was able to discover through simply opening the sample files and looking at the Document Properties. I was not able to open most of them, as they are proprietary and I do not have access to the software.

PDF/A	3D PDF/ U3D	PRC	DXF	DWG	X3D	STEPFile
xmp:CreateDate			Title	Title	title	description
xmp:CreatorTool			Subject	Subject	description	implementation_level
xmp:ModifyDate			Author	Author	created	name
xmp:MetadataDate			Keywords	Keywords	modified	time_stamp
pdf:Producer			Comments	Comments	creator	author
dc:format			Last saved by	Last saved by	Image	organization
dc:title			Revision number	Revision number	reference	preprocessor_version
dc:creator			Create date time	Create date time	identifier	originating_system
xmpMM:DocumentID			Modified date time	Modified date time	license	authorization
xmpMM:InstanceID					generator	
xmpMM:RenditionClass					subject	
xmpMM:VersionID						
xmpMM:History						
pdfaid:part						
pdfaid:conformance						

pdfaExtension:schemas						

Appendix C - Metadata Recommendations – Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.example.com/cadmd">
  <xs:element name="cad">
      <xs:complexType>
          <xs:sequence>
              <xs:element name="PageCount" minOccurs="0" maxOccurs="1"
type="xs:positiveInteger"/>
              <xs:element name="Language" type="xs:string" minOccurs="0"
maxOccurs="unbounded">
                  <xs:annotation>
                      <xs:documentation>
                          A language identifier specifying the natural language used in the
document
                      </xs:documentation>
                  </xs:annotation>
              </xs:element>
              <xs:element name="Font" minOccurs="0" maxOccurs="unbounded">
                  <xs:complexType>
                      <xs:attribute name="name" use="required" type="xs:string"/>
                      <xs:attribute name="isEmbedded" use="optional" type="xs:boolean"/>
                  </xs:complexType>
              </xs:element>
              <xs:element name="Features" minOccurs="0" maxOccurs="unbounded">
                  <xs:simpleType>
                      <xs:restriction base="xs:string">
                          <xs:enumeration value="hasAnnotations">
                              <xs:annotation>
                                  <xs:documentation>
                                      Model includes textual annotations
                                  </xs:documentation>
                              </xs:annotation>
                          </xs:enumeration>
                          <xs:enumeration value="hasKinematics">
                              <xs:annotation>
                                  <xs:documentation>
                                      Model includes kinematics describing the motion of
objects
                                  </xs:documentation>
                              </xs:annotation>
                          </xs:enumeration>
                          <xs:enumeration value="hasMaterialProperties">
                              <xs:annotation>
                                  <xs:documentation>
                                      Includes physical properties of materials used in
model
                                  </xs:documentation>
                              </xs:annotation>
                          </xs:enumeration>
                          <xs:enumeration value="hasTolerances">
                              <xs:annotation>
                                  <xs:documentation>
                                      Measurements within the object are marked with
allowable error tolerances
                                  </xs:documentation>
                              </xs:annotation>
                          </xs:enumeration>
                          <xs:enumeration value="hasTransparencies">
                              <xs:annotation>
                                  <xs:documentation>
                                      Some surfaces or objects within the object are marked
as transparent
                                  </xs:documentation>
                              </xs:annotation>
                          </xs:enumeration>
                      </xs:restriction>
                  </xs:simpleType>
              </xs:element>
              <xs:element name="Representation" minOccurs="0" maxOccurs="unbounded">
                  <xs:complexType>
                      <xs:attribute name="type" use="required">
                          <xs:simpleType>
                              <xs:restriction base="xs:string">
                                  <xs:enumeration value="has2DPointSets"/>
                                  <xs:enumeration value="has3DPointSets"/>
                                  <xs:enumeration value="has2DRasterData">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains two dimensional raster data
or textured surfaces in a three dimensional model
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="has3DRasterData">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains voxel data or stacked 2d
raster data
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="hasImplicitCurves">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains two dimensional lines or
curves defined by implicit equations
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="hasBREP">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains shapes defined by boundary
representation methods
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="hasImplicitSurfaces">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains three dimensional surfaces
defined by implicit equations
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="hasParametricCurves">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains two dimensional lines or
curves defined by parametric equations
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="hasParametricSurfaces">
                                      <xs:annotation>
                                          <xs:documentation>
                                              Object contains three dimensional surfaces
defined by parametric equations
                                          </xs:documentation>
                                      </xs:annotation>
                                  </xs:enumeration>
                                  <xs:enumeration value="hasTriangleMesh"/>
                              </xs:restriction>
                          </xs:simpleType>
                      </xs:attribute>
                      <xs:attribute name="objectCount" use="optional"
type="xs:positiveInteger"/>
                  </xs:complexType>
              </xs:element>
              <xs:element name="Units" minOccurs="0" maxOccurs="unbounded">
                  <xs:complexType>
                      <xs:attribute name="type" use="required">
                          <xs:simpleType>
                              <xs:restriction base="xs:string">
                                  <xs:enumeration value="metric"/>
                                  <xs:enumeration value="imperial"/>
                              </xs:restriction>
                          </xs:simpleType>
                      </xs:attribute>
                  </xs:complexType>
              </xs:element>
              <xs:element name="Extent" minOccurs="0" maxOccurs="1">
                  <xs:complexType>
                      <xs:sequence>
                          <xs:element name="Dimension" minOccurs="1" maxOccurs="3">
                              <xs:complexType>
                                  <xs:attribute name="axis" use="required">
                                      <xs:simpleType>
                                          <xs:restriction base="xs:string">
                                              <xs:enumeration value="x"/>
                                              <xs:enumeration value="y"/>
                                              <xs:enumeration value="z"/>
                                          </xs:restriction>
                                      </xs:simpleType>
                                  </xs:attribute>
                                  <xs:attribute name="magnitude" use="required"
type="xs:decimal"/>
                                  <xs:attribute name="units" use="optional"
type="xs:string"/>
                              </xs:complexType>
                          </xs:element>
                      </xs:sequence>
                  </xs:complexType>
              </xs:element>
          </xs:sequence>
      </xs:complexType>
  </xs:element>
</xs:schema>
```

Appendix D - Metadata Recommendations – Sample Files

X3D NIST/5000points.x3d

```xml
<?xml version="1.0" encoding="UTF-8"?>
<cadmd:cad xmlns:cadmd="http://www.example.com/cadmd" >
  <Representation type="has3DPointSets" objectCount="5000"/>
  <Extent>
      <Dimension axis="x" magnitude="6"/>
      <Dimension axis="y" magnitude="5.4"/>
      <Dimension axis="z" magnitude="6"/>
  </Extent>
</cadmd:cad>
```

3D PDF pdf3d.com/Kompas-Stanchion_eng.pdf

```xml
<?xml version="1.0" encoding="UTF-8"?>
<cadmd:cad xmlns:cadmd="http://www.example.com/cadmd" >
  <PageCount>1</PageCount>
  <Language>English</Language>
  <Font name="Calibri" isEmbedded="true"/>
  <Representation type="hasBREP" objectCount="28"/>
<Extent>
  <Dimension axis="x" magnitude="160"/>
  <Dimension axis="y" magnitude="450"/>
  <Dimension axis="z" magnitude="150"/>
</Extent>
</cadmd:cad>
```

AutoCAD 2015 Mech AutoDesk/Trolley_Structure.dwg

```xml
<?xml version="1.0" encoding="UTF-8"?>
<cadmd:cad xmlns:cadmd="http://www.example.com/cadmd" >
  <Representation type="hasParametricCurves" objectCount="603"/>
  <Extent>
      <Dimension axis="x" magnitude="696.5"/>
      <Dimension axis="y" magnitude="290"/>
  </Extent>
</cadmd:cad>
```