The National Archives

PRONOM Vocabulary Specification: DRAFT

Namespace Document 25 May 2011

This version:

n/a

Latest Version:

n/a

Previous Version:

n/a

Authors:

The National Archives, UK

Contributors:

n/a

Abstract

This specification describes the draft PRONOM vocabulary, defined as a dictionary of named properties and classes using W3C's RDF technology.

Status of this document

The PRONOM vocabulary has been created as part of the Linked Data PRONOM project run by The National Archives. PRONOM is a technical registry that contains information about File-Formats, Software-Packages, Compression-Types and Character-Encodings. The project will convert this information into a linked data format to support information sharing and reuse, to aid with preservation of digital objects over the long term. The work is being run in conjunction with that of data.gov.uk. The namespace resides at http://reference.data.gov.uk, a domain for high quality government reference data and properties.

The National Archives welcome comments on this document, preferably via the project site:

http://labs.nationalarchives.gov.uk/wordpress/index.php/2011/01/linked-data-and-pronom/

The National Archives

Table of Contents

The National Archives

1 PRONOM Vocabulary at a Glance

PRONOM provides a set of classes and properties to aid users in describing file-formats and to allow users to easily describe digital objects using those classes. The vocabulary outlines four classes and a range of other properties to achieve this. Classes and properties are described in the following pages.

1.1 Classes	1.2 Properties
Class: pronom:file-format	Property: pronom:PUID
Class: pronom:compression-type	Property: pronom:XPUID
Class: pronom:character-encoding	Property: pronom:Wave_Format_GUID
Class: pronom:software-package	Property: pronom:MIMETYPE
	Property: pronom:UTI
Class: pronom:Aggregate	Property: pronom:extension
Class: pronom:Audio	Property: pronom:internalSignature
Class: pronom:Database	Property: pronom:lossiness
Class: pronom:GIS	Property: pronom:byteOrder
Class: pronom:Image_(Raster)	Property: pronom:mediaFormat
Class: pronom:Image_(Vector)	Property: pronom:version
Class: pronom:Page_Description	Property: pronom:formatType
Class: pronom:Presentation	Property: pronom:releaseDate
Class: pronom:Spreadsheet	Property: pronom:withdrawnDate
Class: pronom:Text_(Mark-up)	Property: pronom:developedBy
Class: pronom:Text_(Structured)	Property: pronom:supportedBy
Class: pronom:Text_(Unstructured)	
Class: pronom:Video	
Class: pronom:Word_Processor	

2 Who is the PRONOM vocabulary for?

Primarily the vocabulary is for use in the field of digital preservation. The vocabulary allows PRONOM users to build a technical-registry of institution specific information about file formats that they may have in their digital repository, but to allow this information to be shared across the wider internet in the form of linked data.

3 The PRONOM vocabulary description

This specification serves as the PRONOM namespace document. As such it describes the PRONOM vocabulary and the terms (RDF classes and properties) that constitute it, so that Semantic Web applications can use those terms in a variety of RDF-compatible document formats and applications.

3.1 Example

A basic document describing a file-format using the PRONOM vocabulary and a combination of other common vocabulary elements.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ns0="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:ns1="http://reference.data.gov.uk/technical-registry/"
  xmlns:ns2="http://www.w3.org/2004/02/skos/core#"
  xmlns:ns3="http://purl.org/dc/elements/1.1/">

  <rdf:Description rdf:about="http://reference.data.gov.uk/id/file-format/1">

  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdf:type rdf:resource="http://reference.data.gov.uk/technical-registry/file-format"/>
  <ns0:label xml:lang="en">Broadcast WAVE</ns0:label>
  <ns1:version>0</ns1:version>
  <ns2:altLabel xml:lang="en">BWAVE (0)</ns2:altLabel>
  <ns1:formatType rdf:resource="http://reference.data.gov.uk/technical-registry/formatType/Audio"/>
  <ns3:description xml:lang="en">
       Broadcast WAVE is a chunk-based audio format developed by the European Broadcasting Union,
       and based on the Microsoft WAVE format, which is in turn based on the generic Resource
       Interchange File Format (RIFF) specification developed by Microsoft and IBM. Structurally, a
       BWAVE file is composed of a number of chunks, each comprising a four character code chunk
       identifier, the chunk size, and the chunk data. It comprises a RIFF header with a WAVE data
       type identifier, followed by a series of chunks. Every file must include a Broadcast Audio
       Extension chunk, containing metadata required for exchange of information between
       broadcasters, a Format chunk, which describes the format of the audio data, and a Data
chunk,  containing the audio data itself. BWAVE files which contain MPEG-encoded audio data must
also    include a Fact chunk, containing file-dependent information about the audio data, and an
MPEG    Audio Extension chunk, containing extra information required to describe the MPEG encoding.
  </ns3:description>
  <ns1:byteOrder rdf:resource="http://reference.data.gov.uk/technical-registry/Little_endian"/>
  <ns1:releaseDate rdf:datatype="http://www.w3.org/2001/XMLSchemadate">1997-01-01</ns1:releaseDate>
  <ns1:withdrawnDate rdf:datatype="http://www.w3.org/2001/XMLSchemadate">
       2001-07-01
  </ns1:withdrawnDate>
  <ns1:MIMETYPE>audio/x-wav</ns1:MIMETYPE>
  <ns1:PUID>fmt/1</ns1:PUID>
  <ns1:extension>wav</ns1:extension>
  <ns1:internalSignature rdf:resource="http://reference.data.gov.uk/technical-
       registry/internalSignature/129"/>
  <ns1:internalSignature rdf:resource="http://reference.data.gov.uk/technical-
       registry/internalSignature/159"/>

  </rdf:Description>

</rdf:RDF>
```

4 PRONOM vocabulary cross-reference: Listing PRONOM Classes and Properties

The PRONOM vocabulary uses the namespace prefix ‘pronom’ this is a shorthand method to refer to the URL http://reference.data.gov.uk/technical-registry/ - the characters proceeding the namespace prefix are simply concatenated with it, such that pronom:file-format refers to the uri http://reference.data.gov.uk/technical-registry/file-format.

The PRONOM vocabulary introduces the following classes and properties:

4.1 Classes

There are four classes given as follows:

Class: pronom:file-format	
Label	File format
Comment	A file format is an encoded digital object, which may be a file, or a bit stream embedded within a file, and which may be processed or rendered in human or machine readable form. It is an arbitrary method of storing digital content in a file, allowing its later retrieval or interchange with other people and very specific combinations of hardware and software. There are many different formats for different kinds of digital content and often formats can have multiple versions. File formats can be software-independent, or can be developed in conjunction with specific software products.
seeAlso	http://dbpedia.org/resource/File_format
isDefinedBy	pronom: .

Class: pronom:compression-type	
Label	Compression Type
Comment	A method of encoding binary data that reduces its overall size. Compression may be lossless or lossy. Lossless compression stores information more efficiently, but ensures the data is the same decoded as it was before encoding. Lossy compression techniques remove information considered to be irrelevant to the representation, meaning that the file is different when decoded to how it was before encoding. Compression is possible because most digital objects contain superfluous data that is not essential to represent the digital object. This superfluous data is known as redundancy. A digital object may be uncompressed.
isDefinedBy	pronom: .

Class: pronom:character-encoding	
Label	Character Encoding
Comment	The pairing of a representation of characters and symbols with a machine understandable encoding such as hexadecimal digits. ASCII (American Standard Code for Information Interchange) pairs 128 English alphanumeric-characters with the first 128 integers available in the first 7-bits of a byte such that a computer can read these numeric values and understand the character to output.
isDefinedBy	pronom:

Class: pronom:software-package	
Label	Software Package
Comment	Individual programs or a suite of programs that are executed by the computer to accomplish a single task. Software packages exist to perform a wide variety of functions from operating system basic scripting to web development. Software packages require a specific combination of hardware and operating system in order to function.
isDefinedBy	pronom: .

4.2 Properties

Properties are defined as follows:

Property: pronom:PUID	
Label	PUID
Comment	The PRONOM Persistent Unique Identifier (PUID) is an extensible scheme for providing persistent, unique and unambiguous identifiers for records in the PRONOM registry. Such identifiers are fundamental to the exchange and management of digital objects, by allowing human or automated user agents to unambiguously identify, and share that identification of, the representation information required to support access to an object. A PUID is assigned on entry to PRONOM, and every identifier is unique within the namespace, and is persistently and unambiguously bound to a single registry entry. The PUID type element identifies the type of registry entry; for a file format it is fmt or x-fmt; for a character encoding it is chr or x-chr; for a compression encoding it is cmp or x-cmp; for a software package it is sfw or x-sfw. The integer value that follows is the identifier element and denotes an instance of that type within PRONOM.
isDefinedBy	pronom:

Property: pronom:XPUID	
Label	X-PUID
Comment	A legacy form of the PRONOM Persistent Unique Identifier (PUID) originally intended as a temporary identifier to assign to records until a judgment could be made about the validity of the entry; if the entry was deemed to be a valid entry then a standard PUID was meant to replace the X-PUID on an entry. Many collections have been characterised using X-PUIDs and for this reason existing X-PUIDs are still maintained but will not be assigned to any new entries.
isDefinedBy	pronom:

Property: pronom:Wave_Format_GUID	
Label	Wave Format GUID ;
Comment	A globally unique identifier used as a reference for a Waveform Audio File Format. ;
isDefinedBy	pronom: .

The National Archives

Property: pronom:MIMETYPE	
Label	MIME type
Comment	MIME (Multipurpose Internet Mail Extensions) types describe the media type of content either in email or served by web servers or web applications and are intended to help guide a web browser in how the content is to be processed and displayed. Not all formats have an associated MIME type, but this is specified where it is known. MIME is an acronym for Multipurpose Internet Mail Extensions, and is written to allow MIME to be extended in certain ways, without having to revise the standard. MIME types consist of a standardised system of identifiers consisting of a type and a sub-type, separated by a slash, such as text/html. This provides a procedure for maintaining and extending these sets of values by registering them with the Internet Assigned Numbers Authority.
seeAlso:	http://dbpedia.org/resource/MIME
isDefinedBy	pronom:

Property: pronom:UTI	
Label	UTI
Comment	Apple Inc has defined a syntax for data identifiers called uniform type identifiers (UTIs). Each UTI provides a unique identifier for a common system object, such as a particular document or image file type. The UTI is a Unicode text string resource that usually contains characters from a subset of the ASCII character set, such as the Roman alphabet in upper and lower case, the digits 0 through 9, the dot (“.”), and the hyphen (“-”). Colons and slashes are prohibited. It is an extensible system, and Apple and third party developers are able to create UTIs. The UTI type, where it exists, is stored in the metadata of an object.
isDefinedBy	pronom:

Property: pronom:extension	
Label	File Extension
Comment	A file extension is a suffix to the name of a computer file applied to indicate the type of file format. It gives a human readable identifier for the file; such that users can quickly understand the type of file it is without having to open it. The filename extension associates the file with certain software packages, helping an application program recognise whether it is of a type that it can work with. In some operating systems, such as DOS, a file extension is required, but in others, such as Unix, it is optional. Some operating systems limit the length of the extension and some are case sensitive. On Windows computers, extensions consist of a dot '.' at the end of a file name, typically followed by three letters to identify the type of file. On a Unix based system, the file name is a single string, with the '.' being just another character, and with the file name being of variable-length. A file extension is not a reliable identifier of the format of the file. An extension may be linked to more than one program; they are not assigned by a controlling authority, and can be easily changed.
isDefinedBy	pronom:
domain	pronom:file-format

Property: pronom:internalSignature	
Label	Internal Signature
Comment	In PRONOM, we attempt to associate each file format it with a binary signature, termed an internal signature. Internal signatures are created from unique byte sequences that are common within digital objects of the same type. The internal signature is a representation of these byte sequences using hexadecimal notation and simple regular expression features that allow signature developers to create more precise signatures. A signature byte sequence is modelled by describing its starting position within a bitstream and its value. The starting position can be either absolute (the byte sequence starts at a fixed position within the bitstream) or variable (the byte sequence starts at any offset within the bitstream). By definition, a file format specification imposes a specific structure upon the content of the bitstream, which is consistent between all digital objects in that format. The characteristics of this structure may therefore be used as a signature for identifying the format. ;
isDefinedBy	pronom:
domain	pronom:file-format

Property: pronom:lossiness	
Label	Lossiness
Comment	The degree to which data is lost during file compression, which may be given as lossless, lossy, or unknown. Lossless compression stores information more efficiently but ensures the data is the same decoded as it was before encoding. Lossy compression techniques remove information considered to be irrelevant to the representation, meaning that the file is different when decoded to how it was before encoding. Compression is possible because most digital objects contain superfluous data that is not essential to represent the digital object.
isDefinedBy	pronom:
domain	pronom:compression-type

Property: pronom:byteOrder	
Label	Byte order
Comment	Indicates the byte order, or endianness, of binary data which is crucial to its interpretation. The byte order may be little endian, big endian, big and little endian (mixed/middle-endian), or big or little endian (bi-endian). Typically a format specification will state the correct byte-ordering to be used, but it can easily be discovered by analysing the hexadecimal values of the data stream.
isDefinedBy	pronom:

Property: pronom:mediaFormat	
Label	Media format - *[Consider revision to Distribution Medium]*
Comment	The type of storage media on which the software is supplied. Different distribution techniques include CD-ROM, DVD, the internet, and floppy disk. Knowing the distribution medium can help in identifying issues in reading software packages required for digital preservation. Awareness of the physical lifecycle of the format, or other limitations in accessing the content stored on it, are useful indicators of future problems.
isDefinedBy	pronom:

Property: pronom:version	
Label	Version
Comment	The specific version number or letter of the compression technique, file format or character encoding. It is the number or letter used to distinguish this version from previous and subsequent versions, and usually follows the naming convention established by the manufacturer.
isDefinedBy	pronom:

Property: pronom:formatType	
Label	Format Type
Comment	The generic grouping under which the format is classed. This describes the broad content associated with a file, such as whether it is word processed or a database. File formats may have multiple format types, which is represented by the format entry being assigned multiple formatType properties.
isDefinedBy	pronom:
domain	pronom:file-format

Property: pronom:releaseDate	
Label	Release Date
Comment	The date the file format, software package, compression technique or character encoding was released. *[Statement of accuracy here? Flexibility of XSD:Date is questionable]*
isDefinedBy	pronom:

Property: pronom:withdrawnDate	
Label	Withdrawn Date
Comment	The date that support from the compression method creator, file format creator, or support company was withdrawn, or, the date that support for the encoding was withdrawn or superseded. *[Statement of accuracy here? Flexibility of XSD:Date is questionable]*
isDefinedBy	pronom:

Property: pronom:developedBy	
Label	Developed by
Comment	The registered creator of the file format, software product, compression or encoding method.
isDefinedBy	pronom:

Property: pronom:supportedBy	
Label	Supported by
Comment	The registered name of the actor providing post-creation support to the file format, software product, compression or encoding method.
isDefinedBy	pronom:

4.3 Additional Classes: Format Types

The following describes a number of 'Format Types' that PRONOM uses.

Question: Do complications arise when we consider that a file format may be multiple types at the same time?

Question: Should we use nicer URIs? E.g. Image_(Raster) could become pronom:raster-image.

Class: pronom:Aggregate	
Label	Aggregate
Comment	Indicates a format that represents aggregations of arbitrary content drawn from multiple content genre categories. A ZIP file could be a useful example of such an aggregate format, wherein it may contain a wide range and large number of different file format types.
isDefinedBy	pronom:

Class: pronom:Audio	
Label	Audio
Comment	Audio encoded in a digital form which can be decoded and output via a computer. They are digital representations of audio waveforms, and are files that are intended to be heard rather than read. An audio file may be a raw bitstream, but is often presented as part of a multimedia format.
seeAlso	http://dbpedia.org/resource/Audio
isDefinedBy	pronom:

Class: pronom:Database	
Label	Database
Comment	Indicates a format that consists of an organised collection of information in a suitable database management system format and easily accessed by a software application. Binary in nature and characterised by being difficult to open in a text editor, and mostly used by an application specific for the purpose of opening.
seeAlso	http://dbpedia.org/resource/Database
isDefinedBy	pronom:

Class: pronom:GIS	
Label	GIS (Global Information System)
Comment	Indicates a format used within a Geographic Information System. The format may facilitate the capture, storage, manipulation, analysis, display and retrieval of data linked to geographic locations. It enables the linking of geographically referenced data to textual attributes held in a database.
isDefinedBy	pronom:

Class: pronom:Image_(Raster)	
Label	Raster image
Comment	A raster image is a data structure that allows the representation of an image from a series of bits of information. It is intended that the bits are translated into pixels for representation on a display medium. Each pixel is assigned a specific value which determines its colour, and these pixels create an overall finished image.
isDefinedBy	pronom:

Class: pronom:Image_(Vector)	
Label	Vector image
Comment	Images created using geometric primitives such as lines, arcs, and circles and commonly stored using mathematical equations and coordinates. A vector image defines points and the paths that connect them to form a digital representation of an image. Usually contains very little data, but typically includes the starting point (pixel) of the object, the kind of object it is, its size, and colour.
isDefinedBy	pronom:

Class: pronom:Page_Description	
Label	Page Description
Comment	A textual or binary data stream that can be run through an interpreter to generate an image, such as PostScript in creating a PDF (Portable Document Format). As a language it is used to encode documents by precisely describing their appearance when rendered for print or display.
isDefinedBy	pronom: .

Class: pronom:Presentation	
Label	Presentation
Comment	Indicates a format that represents interactive presentation content. Used by a presentation program to edit, manipulate and/or display visual information. A presentation file often contains both text and graphic elements, and is created with the intention of conveying information to a group of people all at once.
seeAlso	http://dbpedia.org/resource/Presentation
isDefinedBy	pronom:

Class: pronom:Spreadsheet	
Label	Spreadsheet
Comment	Indicates a tabular format that allows users to enter numerical data and often manipulate that data into new spreadsheets or graphical representations. Typically uses a software program to record, maintain and display numerical data in rows and columns. Each row and each column is assigned a value, and the intersection of these values is a cell, representing a particular location within the spreadsheet.
seeAlso	http://dbpedia.org/resource/Spreadsheet
isDefinedBy	pronom:

Class: pronom:Text_(Mark-up)	
Label	Mark-up text
Comment	Text which includes a mark-up scheme to be interpreted specifically for other applications. It is used for annotating text in a way that is syntactically distinguishable from that text. Mark-up text is machine-readable and is typically absent from the version of the text which is displayed for end-user consumption.
isDefinedBy	pronom:

Class: pronom:Text_(Structured)	
Label	Structured text
Comment	Simple plain text which is structured for use by other applications. A structured string consists of a sequence of paragraphs separated by one or more blank lines. Each paragraph has a level which is defined as the minimum indentation of the paragraph. A paragraph is a sub-paragraph of another paragraph if the other paragraph is the last preceding paragraph that has a lower level. Structured text might include source code which is to be compiled into an executable format at a later date.
isDefinedBy	pronom:

Class: pronom:Text_(Unstructured)	
Label	Unstructured text
Comment	Unformatted, plain text files that do not fit with a predefined data model, and do not fit well into relational tables. Typically it is text-heavy, and may contain dates and numbers.
isDefinedBy	pronom:

Class: pronom:Video	
Label	Video
Comment	Formats that allow the time-based presentation of still images to give the illusion of movement. A video file may be presented alongside other file types, such as audio content, as part of a multimedia format.
seeAlso	http://dbpedia.org/resource/Video
isDefinedBy	pronom:

Class: pronom:Word_Processor	
Label	Wordprocessed text
Comment	Formatted text and additional content for presentation purposes. Used by a word processor software application to create, edit, format, save and print files. Wordprocessed text may be presented alongside other file types, such as graphics and photographs. Wordprocessed text files typically record presentation information such as character style and size, or other details that specify the appearance of a finished document.
isDefinedBy	pronom:

4.4 Miscellaneous Resources

Resources we've tried to include in PRONOM but would like to assert greater control over how they are described:

Property: pronom:Little_endian	
Label	Little endian
Comment	Endianness refers to a representation method for storing or transmitting binary data. There are four other forms of endianness. Little endian is a representation in which the least significant bit or byte is presented first. An important illustration of endianness is the understanding of integer values. Given the integer, 32,500 in hexadecimal this is represented as 0x7EF4. This is a big-endian ordering. If it is read in hexadecimal with the bytes swapped around, i.e. little-endian, it would be read as 0xF47E which is 62,590 in decimal.
seeAlso	http://dbpedia.org/resource/Endianness
isDefinedBy	pronom:

Property: pronom:Big_endian	
Label	Big endian ;
Comment	Endianness refers to a representation method for storing or transmitting binary data. There are four other forms of endianness. Big endian is a representation in which the most significant bit or byte is presented first. An important illustration of endianness is the understanding of integer values. Given the integer, 32,500 in hexadecimal this is represented as 0x7EF4. This is a big-endian ordering. If it is read in hexadecimal with the bytes swapped around, i.e. little-endian, it would be read as 0xF47E which is 62,590 in decimal.
seeAlso	http://dbpedia.org/resource/Endianness
isDefinedBy	pronom:

5 Glossary

To define the Pronom vocabulary the following other properties and classes have been used:

isDefinedBy

Outlined in the RDF schema vocabulary referenced below rdfs:isDefinedBy points to an original or authoritative description of a resource. The elements of the PRONOM vocabulary are all defined by the resource PRONOM:

domain

Described by the RDF schema vocabulary by asserting a predicate has a domain we can infer that the resource which the predicate is of the type described by the domain. By stating pronom:internalSignature has the domain pronom:file-format we can infer any resource describing a format which uses this property is an instance of pronom:file-format.

Class

Described in more detail in the RDF schema vocabulary we can create a resource describing a class. Using that resource we can then describe other resources. We can say a resource has type file-format. File format is a class; as such the resource is an instance of a file-format.

Property

Described in the RDF concepts document a Property is described as a relation between subject resources and object resources. For example a file format may have a property, endianness.

term_status

Described by the W3C vocabulary status document this property allows us to assert the stability of our vocabulary elements. The terms unstable, testing, stable and archaic allow us to demonstrate our preferred properties and test new properties in a controlled manner.

RDF Concepts document - http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
RDF Schema vocabulary - http://www.w3.org/TR/rdf-schema/
W3C vocabulary status document - http://www.w3.org/2003/06/sw-vocab-status/ns#

6 Acknowledgements

David Tarrant - University of Southampton
John Sheridan - data.gov.uk
Jeni Tennison - TSO
Bill Roberts - Planets advisory board