

PRONOM: Persistenz von PUIDs

Pronom Library Updates und die Persistenz von PUIDs (PRONOM's Persistent Unique Identifier)

Unterschied zwischen x-fmt und fmt-PUIDs

Rund 450 Nummern sind sowohl als fmt als auch als x-fmt in der Pronom Library enthalten. Es gibt keine Relation zwischen z. B. fmt/382 und x-fmt/382 (in diesem Fall Microsoft Visual FoxPro database container und Macromedia FLV). Jeder Identifikator ist eindeutig.

In der Historie waren x-fmt-PUIDs mal als vorübergehende PUIDs gedacht:

PUID types prefixed by 'x-' are used to provide temporary, privat or experimental namespaces for that type. These may be used, for example, where a system requires a PUID identifier to be present which has not yet been formally assigned. Thus, format PUIDs of the type 'x-fmt' might be assigned for formats which have not yet been assigned an 'fmt' identifier. An 'x-' PUID should not be considered persistent. ([PUID Schema von 2006, S. 7](#))

Diese Idee rührt daher, erst eine fmt-PUID zu vergeben, wenn man sicher war, dass die Byte Sequenz (oder eine andere formale Identifizierungsstrategie) korrekt war. Sie wurde aber zugunsten der Persistenz rasch wieder verworfen, die Priorität lautet nun Stabilität der PUIDs ([letzter Paragraph hier](#)).

Deprecated PUIDs

Grundsätzlich bleibt jede PUID persistent, d. h. eine einmal verwendete PUID wird nicht eines Tages für ein anderes Format verwendet. Es kann aber durchaus sein, dass eine PUID fürderhin nicht mehr verwendet wird und somit als "deprecated" gilt. Im Normalfall wird für das Dateiformat dann eine aktuellere PUID verwendet.

Deprecated PUIDs bleiben allerdings auch weiterhin in der Dokumentation erhalten und werden keinesfalls gelöscht (es gibt zwei historische Ausnahmen, x-fmt/366 und x-fmt/431, die zwar gelöscht, aber dokumentiert sind). Die letzte deprecatedPUID ist vom 26.02.2013.

Deprecated PUIDs werden im Feld "Description" des jeweiligen Registryeintrags als Zurückgezogen gekennzeichnet und es wird auf den neuen Identifikator des Formats hingewiesen. Einige Beispiele:

[x-fmt/250](#): PUID deprecated. Please see x-fmt/248 and x-fmt/249 for information on Microsoft Outlook Personal Folders.

[x-fmt/196](#): PUID deprecated. Please see x-fmt/139 for information on NeXT/Sun sound.

[fmt/265](#): PUID deprecated. Please see fmt/163: Microsoft Works Word Processor 1-3 for DOS and 2 for Windows.

[fmt/7](#): PUID deprecated. Please see fmt/353 for information on Tagged Image File Format.

In PRONOM können zurückgezogene PUIDs über eine [Einfache Suche](#) nach "Deprecated" angezeigt werden. Mit akutellem Stand 13.07.2015 sind 58 PUIDs als zurückgezogen gekennzeichnet.

Signature Patterns von deprecated PUIDs anhand des Beispiels TIFF

Während die PUIDs von zurückgezogenen Einträgen weiter bestehen bleiben, werden die Signaturen von zurückgezogenen PUIDs aus dem Registryeintrag entfernt und die Verlinkung zwischen PUID und Pattern in DROID aufgehoben.

Beispiel für das o.g. [fmt/7](#) anhang von DROID Signature Versionen:

File Format Eintrag in DROID Signature Pattern v45 - PUID Status "aktiv":

```
<FileFormat ID="609" MIMEType="image/tiff"
  Name="Tagged Image File Format" PUID="fmt/7" Version="3">
  <InternalSignatureID>9</InternalSignatureID>
  <InternalSignatureID>10</InternalSignatureID>
  <Extension>tif</Extension>
  <Extension>tiff</Extension>
</FileFormat>
```

Das Format [fmt/7](#) ist zur Identifizierung via Signatur den Patternbeschreibungen 9 und 10 zugeordnet. Zur Identifizierung via Formatendung ist [fmt/7](#) tiff und tiff zugeordnet.

File Format Eintrag in DROID Signature Pattern v82 - PUID Status "Deprecated":

```
<FileFormat ID="609" Name="Tagged Image File Format"
  PUID="fmt/7" Version="3"/>
```

Den [Signature Pattern Release Notes](#) ist zu entnehmen, dass fmt/7 mit Pattern Version 51 zurückgezogen wurde. Die Zuordnung von Format zu Signaturen und Formatendungen wurde entfernt. Im konkreten Fall sind die beiden Identifizierungsregeln nun dem "Nachfolge-PUID" fmt/353 zugeordnet:

```
<FileFormat ID="1099" MIMEType="image/tiff"
  Name="Tagged Image File Format" PUID="fmt/353">
  <InternalSignatureID>9</InternalSignatureID>
  <InternalSignatureID>10</InternalSignatureID>
  <Extension>tif</Extension>
  <Extension>tiff</Extension>
</FileFormat>
```

Da auch die ursprünglich für fmt/7 geltenden Pattern 9 und 10 nun für fmt/353 gelten, hat sich an den eigentlichen Signaturpattern nichts geändert und diese bleiben weiter bestehen. Im konkreten Fall der TIFF Signaturen fand lediglich eine Änderung auf dem Level "Specificity" statt. Hierdurch wird gekennzeichnet, ob eine Signatur für mehrere PUIDs gilt ("Generic") oder für nur genau eins ("Specific"). Ferner waren SignatureIDs 9 und 10 vor Version 51 mehreren Formaten zugeordnet (fmt/7, fmt/8, fmt/9, fmt/10) und gelten nunmehr für nur einen PUID (fmt/353) - eine weitere Erläuterung der geänderten TIFF PUIDs befindet sich im nächsten Abschnitt.

```
<InternalSignature ID="9" Specificity="Generic">
  <ByteSequence Reference="B0ffset">
    <SubSequence MinFragLength="0" Position="1"
      SubSeqMaxOffset="0" SubSeqMinOffset="0">
      <Sequence>49492A00</Sequence>
      <DefaultShift>5</DefaultShift>
      <Shift Byte="00">1</Shift>
      <Shift Byte="2A">2</Shift>
      <Shift Byte="49">3</Shift>
    </SubSequence>
  </ByteSequence>
</InternalSignature>
<InternalSignature ID="10" Specificity="Generic">
  <ByteSequence Reference="B0ffset">
    <SubSequence MinFragLength="0" Position="1"
      SubSeqMaxOffset="0" SubSeqMinOffset="0">
      <Sequence>4D4D002A</Sequence>
      <DefaultShift>5</DefaultShift>
      <Shift Byte="00">2</Shift>
      <Shift Byte="2A">1</Shift>
      <Shift Byte="4D">3</Shift>
    </SubSequence>
  </ByteSequence>
</InternalSignature>
```

InternalSignature Eintrag in DROID Signature Pattern 45

```
<InternalSignatureCollect
  <InternalSignature ID
    <ByteSequence Ref
      <SubSequence I
        SubSeqMaxI
          <Sequence:
            <DefaultS
              <Shift By
                <Shift By
                  <Shift By
                    </SubSequence:
  </ByteSequence>
</InternalSignature>
<InternalSignature ID:
  <ByteSequence Ref
    <SubSequence I
      SubSeqMaxI
        <Sequence:
          <DefaultS
            <Shift By
              <Shift By
                <Shift By
                  </SubSequence:
  </ByteSequence>
</InternalSignature>
```

InternalSignature

Neben Änderungen der Zuordnung von Signaturen zu PUIDs sind aber auch bei Signature Pattern selbst Änderungen möglich - diese werden in den DROID Signature Pattern Release Notes angekündigt (z.B. " Macromedia Flash 1. Improved signature through PRONOM research.").

Beispiel deprecated PUID TIFF

So war zwischen 2005 und 2011 die PUID für das TIFF-Format Version 6.0 noch [fmt/10](#) (für die anderen drei TIFF-Versionen fmt/9, fmt/8 und fmt/7). Seit Juli 2011 jedoch wird das TIFF-Format für alle Versionen unter der [fmt/353](#) geführt. Es gibt unter fmt/10 einen Verweis auf fmt/353. Unter der fmt/353 wird zwar angedeutet, dass es Schwierigkeiten mit der Identifikation gab und daher eine neue Interpretation des Standards notwendig war, es gibt einen aber keinen direkten Verweis auf die deprecated PUID fmt/10.

Weiterentwicklung von DROID und Pronom und die praktischen Auswirkungen auf Dateiformatidentifizierung

Aufgrund der Weiterentwicklung der Pronom-Bibliothek wachsen die Möglichkeiten, ein Dateiformat sicher zu erkennen mit jedem Update der Bibliothek. Zum Beispiel war es früher nicht möglich, das Format "epub" zu erkennen, dies wurde lediglich als Containerformat erkannt (auch richtig, aber sehr allgemein).

Außerdem wurden vor der DROID v6 viele Office-Objekte einfach als "OLE2 Compound Document [fmt/11](#)" identifiziert, erst seitdem können sie gezielt als Word 97 -2003 Objekt identifiziert werden. Es ist daher sehr empfehlenswert, im Archiv eine Re- Identifizierung anzustoßen, wenn eine neue Version von DROID bzw. ein Update der Library bereitgestellt wird.

Beispiel vormals nicht erkanntes Format epub

Das [epub-Format](#) wird erst seit 2013 (Droid v6) erkannt. Soweit zurzeit (07/2015) bekannt ist, macht Pronom aber noch keinen Unterschied zwischen epub 2 und epub 3.

Praktischer Einfluss von mittlerweile 82 Pronom Libraries auf den Identifizierungsalltag

Sofern man ein ausreichend großes Sample nimmt und es über die verschiedenen DROID-Versionen / Pronom Libraries laufen lässt, erhält man definitiv für einige Dateien im Laufe der Zeit verschiedene PUIDs.

Die hat diverse Ursachen, z. B.:

- eine bereits existierende PUID war fehlerhaft und wurde korrigiert
- von Zeit zu Zeit werden neue Methoden eingeführt. Z. B. kamen erst mit DROID v6 Container Signatures auf (was auch die Wende bei den Office-Dateien erklärt und auch bei epub (?)).
- oftmals gibt es neue PUIDs und Dateien, die vorher in allgemeinere Kategorien verfrachtet worden waren (siehe [fmt/111](#)). Diese erhielten später dann eine detailliertere Identifizierung.
- manchmal ändert sich auch konzeptionell etwas, so hatte TIFF früher vier PUIDs für die verschiedenen TIFF-Versionen (fmt/7, fmt/8, fmt/9 und fmt/10), seit 2011 ist es aber nur fmt/353 (es gibt noch andere mit dem mimetype/TIFF, aber die Unterteilung in die vier TIFF-Versionen wurde aufgegeben). Hintergrund war, dass es nicht möglich war, die vier TIFF-Versionen verlässlich voneinander zu unterscheiden.

Formatidentifizierung ist keine exakte Wissenschaft. DROID und Pronom - und auch andere Ansätze - werden sich immer weiterentwickeln und somit wird es immer mal wieder zu Änderungen kommen, diese sind ja auch gewünscht und Ergänzungen ja sowieso. Wichtig ist ja vor allem, dass einmal vergebene PUIDs stets weiterhin dokumentiert sind.

Referenzen

[Historische Entwicklung von Pronom](#) (neue PUIDs, Updates usw.)

Fragen kann man auch an die offizielle Mailingliste schicken: PRONOM@nationalarchives.gsi.gov.uk

FAQs Most formats in PRONOM have an 'fmt/' PUID, but some have an 'x-fmt/' PUID. Why is this? <https://groups.google.com/forum/#!topic/droid-list/ZoJwWDjaubQ>