py-pdf / **pypdf** Public

Code Issues 95 Pull requests 28 Discussions Actions Security and

Commit b15a374

stefan6419846 authored last week · 17 / 17 · Verified

```
SEC: Disallow custom XML entity declarations for XMP metadata (#3724)

While *libexpat* already handled the more severe cases, it has still
been possible to cause rather high memory usage. For this reason,
disallow entity declarations completely.

I decided against *defusedxml* for now, as I do not see the benefit of
including an untyped external package for something this small,
especially considering that the public maintenance status does not look
very promising.
```

main (#3724) · 6.10.2 ••• 6.10.0

1 parent d0d9de6 commit b15a374

2 files changed +60 -9 lines changed

Top

Filter files…

- pypdf
  - xmp.py
- tests
  - test_xmp.py

2 files changed +60 -9 lines changed

Search within code

pypdf/xmp.py

```
@@ -15,9 +15,10 @@
15 15         TypeVar,
16 16         Union,
17 17     )
18    - from xml.dom.minidom import Document, parseString
```

```
                 + from xml.dom.expatbuilder import ExpatBuilderNS
                 + from xml.dom.minidom import Document
19               19  from xml.dom.minidom import Element as XmlElement
20               -   from xml.parsers.expat import ExpatError
                 21  + from xml.parsers.expat import ExpatError, XMLParserType
21               22
22               23  from ._protocols import XmpInformationProtocol
23               24  from ._utils import StreamType, deprecate_with_replacement,
                     deprecation_no_replacement
                     @@ -161,6 +162,34 @@ def _generic_get(
161              162         return None
162              163
163              164
                 165 + class _XmpBuilder(ExpatBuilderNS):
                 166 +     """
                 167 +     Custom XML parser denying all entity declarations.
                 168 +
                 169 +     This is a stripped down and typed version inspired by what *defusedxml*
                     does.
                 170 +
                 171 +     Why do we need this? The default limits of *libexpat* used by Python only
                     block exponential entity expansion,
                 172 +     but not cases like quadratic entity expansion which can still cause quite
                     some memory usage.
                 173 +     """
                 174 +
                 175 +     def custom_entity_declaration_handler(
                 176 +             self,
                 177 +             entity_name: str,
                 178 +             is_parameter_entity: bool,
                 179 +             value: Optional[str],
                 180 +             base: Optional[str],
                 181 +             system_id: str,
                 182 +             public_id: Optional[str],
                 183 +             notation_name: Optional[str],
                 184 +     ) -> None:
                 185 +         raise ExpatError(f"Forbidden entities: {entity_name!r}")
                 186 +
                 187 +     def install(self, parser: XMLParserType) -> None:
```

```
      188 +         super().install(parser)
      189 +
      190 +         parser.EntityDeclHandler = self.custom_entity_declaration_handler
      191 +
      192 +
 164  193     class XmpInformation(XmpInformationProtocol, PdfObject):
 165  194         """
 166  195         An object that represents Extensible Metadata Platform (XMP) metadata.
          @@ -175,7 +204,7 @@ def __init__(self, stream: ContentStream) -> None:
 175  204             self.stream = stream
 176  205             try:
 177  206                 data = self.stream.get_data()
 178      -             doc_root: Document = parseString(data)  # noqa: S318
      207 +             doc_root: Document = _XmpBuilder().parseString(data)
 179  208             except (AttributeError, ExpatError) as e:
 180  209                 raise PdfReadError(f"XML in XmpInformation was invalid: {e}")
 181  210             self.rdf_root: XmlElement = doc_root.getElementsByTagNameNS(
```

tests/test_xmp.py

```
          @@ -906,8 +906,11 @@ def
          test_xmp_information__external_entity_expansion(tmpdir):
 906  906       </rdf:RDF>
 907  907     </x:xmpmeta>""".encode())
 908  908
 909      -     xmp = XmpInformation(stream)
 910      -     assert xmp.dc_creator == ["abc"]
      909 +     with pytest.raises(
      910 +             expected_exception=PdfReadError,
      911 +             match=r"^XML in XmpInformation was invalid: Forbidden entities:
          'xxe'$"
      912 +     ):
      913 +         XmpInformation(stream)
 911  914
 912  915
 913  916   @pytest.mark.timeout(10)
          @@ -935,9 +938,28 @@ def
          test_xmp_information__exponential_entity_expansion():
 935  938
 936  939         with pytest.raises(
```

```
937  940                expected_exception=PdfReadError,
938       -             match=(
939       -                 r"^XML in XmpInformation was invalid: limit on input amplification factor "
940       -                 r"\(from DTD and entities\) breached: line 16, column 60$"
941       -             )
     941  +             match=r"^XML in XmpInformation was invalid: Forbidden entities: 'lol'$"
     942  +     ):
     943  +         XmpInformation(stream)
     944  +
     945  +
     946  + @pytest.mark.timeout(10)
     947  + def test_xmp_information__quadratic_entity_expansion():
     948  +     stream = ContentStream(pdf=None, stream=None)
     949  +     stream.set_data(f"""<?xml version="1.0"?>
     950  + <!DOCTYPE lolz [
     951  +   <!ENTITY a "{'A' * 10_000}">
     952  + ]>
     953  + <x:xmpmeta xmlns:x="adobe:ns:meta/">
     954  +   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
     955  +     <rdf:Description rdf:about="">
     956  +       <dc:title xmlns:dc="http://purl.org/dc/elements/1.1/">{'&a;' * 99}</dc:title>
     957  +     </rdf:Description>
     958  +   </rdf:RDF>
     959  + </x:xmpmeta>""".encode())
     960  +
     961  +     with pytest.raises(
     962  +             expected_exception=PdfReadError,
     963  +             match=r"^XML in XmpInformation was invalid: Forbidden entities: 'a'$"
942  964          ):
943  965              XmpInformation(stream)
```

Comments 0

Please sign in to comment.