

Internet Engineering Task Force (IETF)
Request for Comments: 6129
Category: Informational
ISSN: 2070-1721

L. Romary
TEI Consortium and INRIA
S. Lundberg
The Royal Library, Copenhagen
February 2011

The 'application/tei+xml' Media Type

Abstract

This document defines the 'application/tei+xml' media type for markup languages defined in accordance with the Text Encoding and Interchange guidelines.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at <http://www.rfc-editor.org/info/rfc6129>.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

- 1. Introduction 2
- 2. Recognizing TEI Files 2
- 3. Fragment Identifier 4
- 4. Security Considerations 4
 - 4.1. Harmful Content 4
 - 4.2. Intellectual Property Rights 4
 - 4.3. Authenticity and confidentiality 5
- 5. IANA Considerations 5
 - 5.1. Registration of MIME Type 'application/tei+xml' 5
- 6. References 6
 - 6.1. Normative References 6
 - 6.2. Informative References 7

1. Introduction

Text Encoding and Interchange (TEI) is an international and interdisciplinary standard that is widely used by libraries, museums, publishers, and individual scholars to represent all kinds of textual material for online research and teaching [TEI].

This document defines the 'application/tei+xml' media type in accordance with [RFC3023] in order to enable generic processing of such documents on the Internet using eXtensible Markup Language (XML) [W3C.REC-xml-20081126] technologies.

2. Recognizing TEI Files

TEI files are XML documents or fragments having the root element (as defined in [W3C.REC-xml-20081126]) in a TEI namespace. TEI namespace names are defined as a Universal Resource Identifier (URI) [RFC3986] in accordance with [W3C.REC-xml-names-20091208] and begins with <http://www.tei-c.org/ns/> followed by the version number of the namespace. The current namespace is <http://www.tei-c.org/ns/1.0>

The most common root element names for TEI documents are

```
<TEI>

<teiCorpus>
```

The teiCorpus documents provide the ability to bundle multiple documents into a single file.

Examples:

A document having <TEI> root element

```
<?xml version="1.0" encoding="UTF-8" ?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    ...
  </teiHeader>
  <text>
    ...
  </text>
</TEI>
```

A document having <teiCorpus> root element

```
<?xml version="1.0" encoding="UTF-8" ?>
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    ...
  </teiHeader>
  <TEI>
    <teiHeader>
      ...
    </teiHeader>
    <text>
      ...
    </text>
  </TEI>
  <TEI>
    ... second document ...
  </TEI>
  <TEI>
    ... third document ...
  </TEI>
</teiCorpus>
```

TEI and teiCorpus files are often given the extensions .tei and .teiCorpus, respectively. There is a third type of file, which often is given the suffix .odd. ODD ("One Document Does it All") is a TEI XML document that includes schema fragments, prose documentation, and reference documentation. It is used for the definition and documentation of XML-based languages, and primarily for the TEI Guidelines [ODD]. In other words, ODD files do not differ from other TEI files in syntax, only in function.

3. Fragment Identifier

Documents having the media type 'application/tei+xml' use the fragment identifier notation as specified in [RFC3023] for the media type 'application/xml'.

4. Security Considerations

An XML resource does not in itself compromise data security. When being available on a network simply through the dereferencing of an Internationalized Resource Identifier (IRI) [RFC3987] or a URI, care must be taken to properly interpret the data to prevent unintended access. Hence the security issues of [RFC3986], Section 7, apply. In addition, as this media type uses the "+xml" convention, it shares the same security considerations as described in RFC 3023 [RFC3023], Section 10. In general, security issues related to the use of XML in IETF protocols are treated in RFC 3470 [RFC3470], Section 7. We will not try to duplicate this material, but review some aspects that are important for document-centric XML as applied to text encoding.

4.1. Harmful Content

Any application accepting submitted or retrieving TEI XML for processing has to be aware of risks connected with injection of harmful scripts and executable XML. XML inclusion [W3C.REC-xinclude-20061115] and the use of external entities are vulnerable to various forms of spoofing, and can also reveal aspects of a service in a way that may compromise its security. Any vulnerability of these kinds are, however, application specific. The TEI namespaces do not contain such elements.

4.2. Intellectual Property Rights

TEI documents often arise in digitization of cultural heritage materials. Texts made accessible in TEI format may be unrestricted in the sense that their distribution may be unlimited by Digital Rights Management [DRM] or Intellectual Property Rights [IPR] constraints. However, TEI documents are heterogeneous. Some parts of a document may be unrestricted, whereas others, such as editorial text and annotations, may be subject to DRM restrictions.

The TEI format provides means for highly granular attribution, down to the content of individual XML elements. Software agents participating in the exchange or processing TEI may be required to honour markup of this kind. Even when there are no IPR constraints, intellectual property attribution alone requires that document users be able to tell the difference between content from different sources.

4.3. Authenticity and confidentiality

Historical archival records are often encoded in TEI and legal document may be binding centuries after they were written. Digitization and encoding of legal texts may require technologies for assuring authenticity, such as cryptographic checksums and electronic signatures.

Similarly, historical documents may in part or in their entirety be confidential. This may be required by law or by the terms and conditions, such as in the case of donated or deposited text from private sources. A text archive may need content filtering or cryptographic technologies to meet such requirements.

5. IANA Considerations

5.1. Registration of MIME Type 'application/tei+xml'

MIME media type name: application

MIME subtype name: tei+xml

Required parameters: None

Optional parameters: charset

the parameter has identical semantics to the charset parameter of the "application/xml" media type as specified in RFC 3023 [RFC3023].

Encoding considerations:

Identical to those for 'application/xml'. See RFC 3023 [RFC3023], Section 3.2.

Security considerations:

See Security Considerations (Section 4) in this specification.

Interoperability considerations:

TEI documents are often given the extension '.xml', which is not uncommon for other XML document formats.

Published specification:

This media type registration is for TEI documents [TEI] as described here. TEI syntax is defined in a schema [TEISchema].

Applications which use this media type:

There are currently no known applications using the media type 'application/tei+xml'.

Additional information:

Magic number(s):

There is no single initial octet sequence that is always present in TEI documents.

file extension(s):

Common extensions are '.tei', '.teiCorpus' and '.odd'. See Recognizing TEI files (Section 2) in this specification.

Macintosh File Type Code(s)

TEXT

Object Identifier(s) or OID(s)

Not applicable

6. References

6.1. Normative References

- [RFC3023] Murata, M., St. Laurent, S., and D. Kohn, "XML Media Types", RFC 3023, January 2001.
- [RFC3470] Hollenbeck, S., Rose, M., and L. Masinter, "Guidelines for the Use of Extensible Markup Language (XML) within IETF Protocols", BCP 70, RFC 3470, January 2003.
- [RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.
- [RFC3987] Duerst, M. and M. Suignard, "Internationalized Resource Identifiers (IRIs)", RFC 3987, January 2005.
- [TEI] "TEI Guidelines", <<http://www.tei-c.org/Vault/P5/1.8.0/doc/tei-p5-doc/en/html/>>.

[TEISchema]

"Schema generated from ODD source", <http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng>.

[W3C.REC-xml-20081126]

Paoli, J., Yergeau, F., Sperberg-McQueen, C., Maler, E., and T. Bray, "Extensible Markup Language (XML) 1.0 (Fifth Edition)", World Wide Web Consortium Recommendation REC-xml-20081126, November 2008, <<http://www.w3.org/TR/2008/REC-xml-20081126>>.

[W3C.REC-xml-names-20091208]

Bray, T., Hollander, D., Layman, A., Tobin, R., and H. Thompson, "Namespaces in XML 1.0 (Third Edition)", World Wide Web Consortium Recommendation REC-xml-names-20091208, December 2009, <<http://www.w3.org/TR/2009/REC-xml-names-20091208>>.

6.2. Informative References

[DRM]

"Digital rights management", <http://en.wikipedia.org/w/index.php?title=Digital_rights_management&oldid=412653591>.

[IPR]

"Intellectual property", <http://en.wikipedia.org/w/index.php?title=Intellectual_property&oldid=411690322>.

[ODD]

"Getting Started with P5 ODDs", <<http://www.tei-c.org/Guidelines/Customization/odds.xml>>.

[W3C.REC-xinclude-20061115]

Marsh, J., Orchard, D., and D. Veillard, "XML Inclusions (XInclude) Version 1.0 (Second Edition)", World Wide Web Consortium Recommendation REC-xinclude-20061115, November 2006, <<http://www.w3.org/TR/2006/REC-xinclude-20061115>>.

Authors' Addresses

Laurent Romary
TEI Consortium and INRIA

E-Mail: laurent.romary@inria.fr
URI: <http://www.tei-c.org/>

Sigfrid Lundberg
The Royal Library, Copenhagen
Postbox 2149
1016 Koebenhavn K
Denmark

E-Mail: slu@kb.dk
URI: <http://sigfrid-lundberg.se/>

