This document proposes new TEI-related file formats for the PRONOM file format database (http://www.nationalarchives.gov.uk/PRONOM).

I'm not sure whether TEI should get four PUIDs (one for each variant that can be reliably identified) or five (those plus an over-arching one which includes the SGML files too). An example of an SGML file is http://lcweb2.loc.gov/ndlpcoop/nicmoas/oldg/oldg0004.sgm

I've not converted signatures to byte sequences, I've left it in a human-readable form to enable easier peer review by colleagues..Conversion to byte sequences should be as per UTF8.

Thanks to James Cummings, Martin Holmes and Lou Burnard for their valuable feedback.

Organisation name	TEI Consortia
Name	Text Encoding Initiative
Address	
Address country	
Telephone	
Support website	
Company website	http://www.tei-c.org/index.xml
Contact	TEI-L@LISTSERV.BROWN.EDU
Source	Stuart Yeates
Source date	May 2020
Source description	The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. (from http://www.tei-c.org/index.xml) The TEI community provides best-effort support for all TEI matters through the TEI-L mailing list at TEI-L@LISTSERV.BROWN.EDU

P5 text

Name	P5 TEI / XML Single Text
Version	P5
Other names	
Identifiers	MIME: application/tei+xml
Family	Text Encoding Initiative
Classification	Text (Mark-up)
Disclosure	
Description	The Text Encoding Initiative Guidelines (TEI) provide a methodology for encoding textual content for a wide variety of academic and publishing purposes and repurposes. TEI P5 is serialised as XML. Conceptually, TEI is a sibling of HTML which focuses on textual semantics rather than display. The 'ODD' file extension is used when TEI is being used as a literate programming language for XML schemas, modern TEI guidelines are written in this form of the language, HTML, RNG, RNC and DTD files are then generated from the TEI. Note that TEI permits customisations and some may mean that the file no longer matches the signatures. This is especially the case where TEI is fragmented, embedded within other XML or when other XML namespaces are used. TEI customisation is primarily done via TEI-supplied tools, formerly PizzaChef https://tei-c.org/Vault/P4/pizza.html and now Roma https://roma2.tei-c.org/ `P5 differs from the preceding P4 version, amongst other things, using the then-new xml:lang` and `xml:id tags.`
Orientation	
Byte order	
Related file formats	

Technical Environment	
Released	2007
Supported until	
Format Risk	
Developed by	TEI Consortia
Supported by	TEI Consortia
Source	Stuart Yeates
Source date	May 2020
Source description	
Last updated	May 2020
Note	
Documentation	http://www.tei-c.org/Vault/P5/current/doc/tei-p5-doc/en/html/ https://en.wikipedia.org/wiki/Text_Encoding_Initiative
Reference files	Schemas: http://www.tei-c.org/Vault/P5/current/xml/tei/schema/relaxng/tei.rnc http://www.tei-c.org/Vault/P5/current/xml/tei/schema/relaxng/tei.rng http://www.tei-c.org/Vault/P5/current/xml/tei/schema/dtd/ Sample files: https://raw.githubusercontent.com/srophe/places/master/xml/1005.xml https://raw.githubusercontent.com/iulibdcs/tei_text/master/vwwp_tei/VAB7010.xml http://www.ibsen.uio.no/DRVIT_Vi%7CVi4262III1.xml http://purl.dlib.indiana.edu/iudl/law/brevier/encodedtext/VAA8558-01 http://tei.oucs.ox.ac.uk/Talks/2011-02-aix/talk-hdr.xml?style=raw http://ota.ox.ac.uk/text/5268.xml http://buddhistinformatics.ddbc.edu.tw/BZA/getsource.xql?src=bza001eng.xml https://github.com/tei-fr/formationEnc2017-10/blob/master/Travaux/oulipoEd/610417.xml https://sourceforge.net/p/xdge/code/HEAD/tree/xdge/xml/xdge1.xml?format=raw

External signatures	File extension: xml File extension: tei File extension: odd	
Internal signatures	Name	Text Encoding Initiative / XML P5 Text
	Description	*<?xml version="1.0"*TEI*http://www.tei-c.org/ns/1.0
	Byte sequences	

Position type	Absolute from BOF
Offset	0
Maximum Offset	1024
Byte order	
Value	

P5 corpus

Name	P5 TEI / XML - Corpus
Version	P5
Other names	
Identifiers	MIME: application/tei+xml
Family	Text Encoding Initiative
Classification	Text (Mark-up)
Disclosure	
Description	The Text Encoding Initiative Guidelines provide a methodology for encoding textual content for a wide variety of academic and publishing purposes and repurposes. P5 is based on XML. Conceptually, TEI is a sibling of HTML which focuses on textual semantics rather than display. Note that TEI permits customisations and some may mean that the file no longer matches the attached signatures. This is especially the case where TEI is fragmented, embedded within other XML or when other XML namespaces are used. TEI customisation is primarily done via TEI-supplied tools, formerly PizzaChef https://tei-c.org/Vault/P4/pizza.html and now Roma https://roma2.tei-c.org/ `P5 differs from the preceding P4 version, amongst other things, using the then-new xml:lang` and `xml:id tags.`
Orientation	
Byte order	
Related file formats	
Technical Environment	
Released	`2007`
Supported until	

Format Risk	
Developed by	TEI Consortia
Supported by	TEI Consortia
Source	Stuart Yeates
Source date	May 2020
Source description	
Last updated	May 2020
Note	
Documentation	http://www.tei-c.org/Vault/P5/current/doc/tei-p5-doc/en/html/ https://en.wikipedia.org/wiki/Text_Encoding_Initiative
Reference files	Schemas: http://www.tei-c.org/Vault/P5/current/xml/tei/schema/relaxng/tei.rnc http://www.tei-c.org/Vault/P5/current/xml/tei/schema/relaxng/tei.rng http://www.tei-c.org/Vault/P5/current/xml/tei/schema/dtd/ Sample files: http://nlp.ipipan.waw.pl/TEI4NKJP/example_all_levels_1M/text.xml http://www.18thpress.ca/node/514/tei https://github.com/tei-fr/formationEnc2017-10/blob/master/Travaux/oulipoEd/oulipoCorpus.xml

External signatures	File extension: xml File extension: tei File extension: teiCorpus	
Internal signatures	Name	Text Encoding Initiative / XML P5 Corpus

	Description	*<?xml version="1.0"*teiCorpus*http://www.tei-c.org/ns/1.0
	Byte sequences	
		Position type: Absolute from BOF
		Offset: 0
		Maximum Offset: 1024
		Byte order:
		Value:

P4 text

Name	P4 TEI / XML - Single Text
Version	P4
Other names	
Identifiers	MIME: application/tei+xml
Family	Text Encoding Initiative
Classification	Text (Mark-up)
Disclosure	

Description	The Text Encoding Initiative Guidelines provide a methodology for encoding textual content for a wide variety of academic and publishing purposes and repurposes. P4 is serialised as XML. Conceptually, TEI is a sibling of HTML which focuses on textual semantics rather than display. The ‘ODD’ file extension is used when TEI is being used as a literate programming language for XML schemas, modern TEI guidelines are written in this form of the language, HTML, RNG, RNC and DTD files are then generated from the TEI. Note that TEI permits customisations and some may mean that the file no longer matches the attached signatures. This is especially the case where TEI is fragmented, embedded within other XML or when other XML namespaces are used. Information on how to convert TEI P4 to TEI P5 can be found at https://www.tei-c.org/Vault/P4/migrate.html and a script at https://www.tei-c.org/Vault/P4/p4top5.xsl Such automated conversion is not guaranteed to be lossless, particularly when the TEI has been customised. `P4 differs from the proceeding P3 version by being in XML rather than in SGML.`
Orientation	
Byte order	
Related file formats	
Technical Environment	
Released	2001
Supported until	
Format Risk	
Developed by	TEI Consortia
Supported by	TEI Consortia
Source	Stuart Yeates
Source date	May 2020
Source description	

Last updated	May 2020
Note	
Documentation	http://www.tei-c.org/Vault/P4/doc/html/index.html https://en.wikipedia.org/wiki/Text_Encoding_Initiative
Reference files	Sample files: http://docsouth.unc.edu/fpn/andrews/andrews.xml http://inslib.kcl.ac.uk/irt2009/IRT101.xml http://emblems.let.uu.nl/xml/he1601.xml https://raw.githubusercontent.com/ericleasemorgan/epub/master/dracula.xml https://scancan.net/xml/crocker_1_27.xm

External signatures	File extension: xml File extension: tei File extension: odd	
Internal signatures	Name	Text Encoding Initiative / XML P4
	Description	*<?xml version="1.0"*TEI\.2
	Byte sequences	Position type: Absolute from BOF Offset: 0 Maximum Offset: 2048 Byte order: Value:

P4 Corpus

Name	P4 TEI / XML - Corpus
Version	P4
Other names	
Identifiers	MIME: application/tei+xml
Family	Text Encoding Initiative
Classification	Text (Mark-up)
Disclosure	
Description	The Text Encoding Initiative Guidelines provide a methodology for encoding textual content for a wide variety of academic and publishing purposes and repurposes. P4 is serialised as XML. Conceptually, TEI is a sibling of HTML which focuses on textual semantics rather than display. Note that TEI permits customisations and some may mean that the file no longer matches the attached signatures. This is especially the case where TEI is fragmented, embedded within other XML or when other XML namespaces are used. Information on how to convert TEI P4 to TEI P5 can be found at https://www.tei-c.org/Vault/P4/migrate.html and a script at https://www.tei-c.org/Vault/P4/p4top5.xsl Such automated conversion is not guaranteed to be lossless, particularly when the TEI has been customised. `P4 differs from the proceeding P3 version by being in XML rather than in SGML.`

Orientation	
Byte order	
Related file formats	
Technical Environment	
Released	2001
Supported until	
Format Risk	
Developed by	TEI Consortia
Supported by	TEI Consortia
Source	Stuart Yeates
Source date	May 2020
Source description	
Last updated	May 2020
Note	
Documentation	http://www.tei-c.org/Vault/P4/doc/html/index.html https://en.wikipedia.org/wiki/Text_Encoding_Initiative
Reference files	https://raw.githubusercontent.com/cdli-gh/Semantic-Role-Labeler/master/data/etcsl/2518/etcsl/etcsl.xml https://raw.githubusercontent.com/SabrinaWSY/projet_XML_exos/master/Exos9/AF_corpus_070703.xml http://www.natcorp.ox.ac.uk/archive/vault/bncx32.txt

File extension: xml

External signatures	File extension: tei File extension: teiCorpus		
Internal signatures	Name	Text Encoding Initiative / XML P4 Corpus	
	Description	*<?xml version="1.0"*teiCorpus\.2	
	Byte sequences		
		Position type	Absolute from BOF
		Offset	0
		Maximum Offset	2048
		Byte order	
		Value	