

Sunoikisis Digital Classics 2025

Preparing Texts & Data Cleaning

Jonathan Blaney (Cambridge Digital Humanities)

Gabriel Bodard (University of London)

Katharine Shields (King's College London)



Outline

- Why do we need to clean texts?
- What are Regular expressions?
- Regex syntax
- Regex demos
- Python demo
- When not to use regex

Why do we need to clean texts?

Text cleaning

- Removing “noise” from a text for analysis
 - Correcting errors, e.g. typographical errors, OCR artefacts
 - Removing metadata, e.g. line numbers or page numbers
- Significant part of any project working on digitising/digitised texts
- Clean data is task-specific



(j=η sz=š s,=ş t,=ţ 0-9=0-9; 'alef)



< [ITEM 1 of 73]

en

only

Amel-Marduk [6]
 Nabonidus [73]
 Nabopolassar [15]
 Nebuchadnezzar II [130]
 Neriglissar [8]

Names

- Nabonidus 01

Numbers

- Q005398
- Nabonidus 01

View

- Catalogue entry
- Print text de
- Cuneified
- TEI

Details

- cylinder
- Neo-Babylonian
- Babylon
- Royal Inscription
- Nabonidus

Sources

Nabonidus 01

- i 1 ^dAG-na- 'i-id LUGAL TIN.TIR.KI NUN na-a-du
 i 2 re-é-a-am za-ni-nu šá a-na ʔe₄-em DINGIR.MEŠ pu-tuq-qu
 i 3 e-em-qá mu-ut-né-en-nu-ú mu-uš-te-né-e'-ú aš-ra-a-ti
 DINGIR.MEŠ GAL.MEŠ
 i 4 eṭ-lu šu-us-su-mu bi-nu-ut ABGAL DINGIR.MEŠ ^dAMAR.UTU
 i 5 na-ab-ni-it ^dE₄.RU₆ ba-na-a-ta gi-mir ma-al-ku
 i 6 i-ti-it ^dMU.Ú.A.TI a-pil é-sag-íl sa-niq mit-ḥur-tú¹
 i 7 bi-nu-tu ^dnin-ši-kù mu-du-ú ba-nu-ú ka-la-ma
 i 8 ni-bi-it ^dŠEŠ.KI-ri be-lu a-gi-i mu-kal-li-im ša-ad-du
 i 9 ša u₄-mi-šam-ma iš-te-né-e'-ú pu-luḥ-tu₄ DINGIR.MEŠ
 GAL.MEŠ
 i 10 a-na zi-in-na-a-ti é-sag-íl ù é-zi-da
 i 11 ba-ša-a uz-na-a-šu
 i 12 DUMU ^{md}na-bi-um-ba-lat-su-iq-bi NUN e-em-qá a-na-ku
 i 13 URU KÁ.DINGIR.RA.KI a-na dam-qa-a-ti aš-te-né-e'-e
 i 14 a-na é-sag-íl É.GAL DINGIR.MEŠ GAL.MEŠ šu-tú-ra-ku
 IGI.SÁ-e
 i 15 a-na é-zi-da šá-ad ba-la-ṭu mim-ma šum-šu du-uš-šá-ku
 i 16 šá é-mes-lam É qar-ra-du DINGIR.MEŠ tú-uḥ-ḥu-da-ak ḥi-iš-
 bi
 i 17 i-nu-šu im-gur-^dEN.LÍL BÀD KÁ.DINGIR.RA.KI
 i 18 iš-da-a-šu i-nu-šu-ma i-qu-pu i-ga-ru-šu²
 i 19 re-e-ši-šu it-ru-ur-ma né-mé-et-ta la i-ši

(i 1) Nabonidus, king of Babylon, attentive prince, the shepherd who provides, the one who is constantly attentive to the will of the gods, the wise (and) pious one, the one who constantly seeks out the shrines of the great gods, most befitting warrior, creation of the sage of the gods — the god Marduk — product of the goddess Erua — creator of all rulers — selected by the god Nabû — the heir of Esagil who controls (cosmic) harmony — creation of the god Ninšiku — the (all-)knowing creator of everything — chosen by the god Nannāru — the lord of the crown who makes astrological signs known — the one who strives every day (to show) devotion to the great gods (and) whose mind is focused on provisioning Esagil and Ezida, son of Nabû-balāssu-iqbi, wise prince, am I.

(i 13) I constantly seek out the city of Babylon for good deeds. I increase gifts to Esagil, the palace of the great gods; I abundantly supply everything to Ezida, the mountain of life; (and) I lavishly provide abundance to Emeslam, the temple of the hero of the gods.

(i 17) At that time, (with regard to) Imgur-Enlil, the wall of Babylon, its foundations had become shaky, its walls had buckled, its superstructure was tottering, and it had no support.

- i 1 ^dAG-na- 'i-id LUGAL TIN.TIR.KI NUN na-a-du
i 2 re-é-a-am za-ni-nu šá a-na ʔe₄-em DINGIR.MEŠ pu-tuq-qu
i 3 e-em-qá mu-ut-né-en-nu-ú mu-uš-te-né-e'-ú aš-ra-a-tì
DINGIR.MEŠ GAL.MEŠ
i 4 eṭ-lu šu-us-su-mu bi-nu-ut ABGAL DINGIR.MEŠ ^dAMAR.UTU
i 5 na-ab-ni-it ^dE₄.RU₆ ba-na-a-ta gi-mir ma-al-ku
i 6 i-ti-it ^dMU.Ú.A.TI a-pil é-sag-íl sa-niq mit-ḥur-ti **1**
i 7 bi-nu-tu ^dnin-ši-kù mu-du-ú ba-nu-ú ka-la-ma
i 8 ni-bi-it ^dŠEŠ.KI-ri be-lu a-gi-i mu-kal-li-im ša-ad-du
i 9 ša u₄-mi-šam-ma iš-te-né-e'-ú pu-luḥ-tu₄ DINGIR.MEŠ
GAL.MEŠ
i 10 a-na zi-in-na-a-ti é-sag-íl ù é-zi-da
i 11 ba-ša-a uz-na-a-šu
i 12 DUMU ^{md}na-bi-um-ba-lat-su-iq-bi NUN e-em-qá a-na-ku
i 13 URU KÁ.DINGIR.RA.KI a-na dam-qa-a-ti aš-te-né-e'-e
i 14 a-na é-sag-íl É.GAL DINGIR.MEŠ GAL.MEŠ šu-tú-ra-ku
IGI.SÁ-e
i 15 a-na é-zi-da šá-ad ba-la-ṭu mim-ma šum-šu du-uš-šá-ku
i 16 šá é-mes-lam Ē qar-ra-du DINGIR.MEŠ tú-uḥ-ḥu-da-ak ḥi-iš-
bi
i 17 i-nu-šu im-gur-^dEN.LÍL BÀD KÁ.DINGIR.RA.KI
i 18 iš-da-a-šu i-nu-šu-ma i-qu-pu i-ga-ru-šu **2**
i 19 re-e-ši-šu it-ru-ur-ma né-mé-et-ta la i-ši

i 1 |

dAG-na-^pi-id LUGAL TIN.TIR.KI NUN na-a-du

(i 1) Nabonidus, king of Babylon, attentive prince, the shepherd who provides, the one who is constantly att

i 2

re-é-a-am za-ni-nu šá a-na ʔe₄-em DINGIR.MEŠ pu-tuq-qu

i 3

e-em-qá mu-ut-né-en-nu-ú mu-uš-te-né-e^p-ú aš-ra-a-tì DINGIR.MEŠ GAL.MEŠ

i 4

eṭ-lu šu-us-su-mu bi-nu-ut ABGAL DINGIR.MEŠ dAMAR.UTU

i 5

na-ab-ni-it dE₄.RU₆ ba-na-a-ta gi-mir ma-al-ku

i 6

i-ti-it dMU.Ú.A.TI a-pil é-sag-íl sa-niq mit-ḥur-tú1

i 7

bi-nu-tu dnin-ši-kù mu-du-ú ba-nu-ú ka-la-ma

i 8

ni-bi-it dŠEŠ.KI-ri be-lu a-gi-i mu-kal-li-im ša-ad-du

i 9

ša u₄-mi-šam-ma iš-te-né-e^p-ú pu-luḥ-tu₄ DINGIR.MEŠ GAL.MEŠ

i 10

a-na zi-in-na-a-ti é-sag-íl ù é-zi-da

i 11

ba-ša-a uz-na-a-šu

i 12

DUMU mdna-bi-um-ba-lat-su-iq-bi NUN e-em-qá a-na-ku

i 13

URU KÁ.DINGIR.RA.KI a-na dam-qa-a-ti aš-te-né-e^p-e

(i 13) I constantly seek out the city of Babylon for good deeds. I increase gifts to Esagil, the palace of t

i 14

RINBE 2, Nabonidus 01 composite (P518907)

Official or display artifact excavated in Uncertain (mod. uncertain), dated to the Neo-Babylonian (ca. 626-539 BC) period

[Export artifact](#) ▾[History](#) ▾

Metadata / catalogue

[Flat catalogue](#)[As CSV](#)[As TSV](#)[Expanded catalogue](#)[As JSON](#)[Linked catalogue](#)[As TTL](#)[As JSON-LD](#)[As RDF/JSON](#)[As RDF/XML](#)

Chemical Data

[Seal Chemistry](#)[As CSV](#)

Text / annotations

[Text data](#)[As ATF](#)[As JTF](#)[Linked annotations](#)[As TTL](#)[As JSON-LD](#)[As RDF/JSON](#)[As RDF/XML](#)

Related publications

[As CSV](#)[As BibTeX](#)

NUMBER

ENCE
(mod. uncertain)

L(S)

SE(S)

SITE

[Expand All](#)[Collapse all](#)

Text



&P518907 = RINBE 2, Nabonidus 01 composite

#atf: lang akk

@object composite text

@surface a

1. {d}na3-na-'i-id _lugal_ babilax(|TIN.TIR|){ki} _nun_ na-a-du

#tr.en: Nabonidus, king of Babylon, attentive prince,

2. re-e2-a-am za-ni-nu sza2 a-na t,e4-em _dingir-mesz_ pu-tuq-qu

#tr.en: the shepherd who provides, the one who is constantly attentive to the will of the gods,

3. e-em-qa2 mu-ut-ne2-en-nu-u2 mu-usz-te-ne2-e'-u2 asz-ra-a-ti3 _dingir-mesz_ gal-mesz_

#tr.en: the wise (and) pious one, the one who constantly seeks out the shrines of the great gods,

4. et,-lu szu-us-su-mu bi-nu-ut _abgal dingir-mesz_ {d}marduk

#tr.en: most befitting warrior, creation of the sage of the gods—the god Marduk—

5. na-ab-ni-it {d}e4-ru6 ba-na-a-ta gi-mir ma-al-ku

#tr.en: product of the goddess Erua—creator of all rulers—

6. i-ti-it {d}mu-u2-a-ti a-pil e2-sag-il2 sa-niq mit-hur-tu2

#tr.en: selected by the god Nabû—the heir of Esagil who controls (cosmic) harmony—

7. bi-nu-tu {d}nin-szi-ku3 mu-du-u2 ba-nu-u2 ka-la-ma

#tr.en: creation of the god Ninšiku—the (all-)knowing creator of everything—

8. ni-bi-it {d}nanna-ri be-lu a-gi-i mu-kal-li-im s,a-ad-du

#tr.en: chosen by the god Nannāru—the lord of the crown who makes astrological signs known—

9. sza u4-mi-szam-ma isz-te-ne2-e'-u2 pu-luh-tu4 _dingir-mesz_ gal-mesz_

#tr.en: the one who strives every day (to show) devotion to the great gods

10. a-na zi-in-na-a-ti e2-sag-il2 u3 e2-zi-da

#tr.en: (and) on provisioning Esagil and Ezida,

&P269975 = RINBE 2, Nabonidus 01, ex. 01

#atf: lang akk

@object barrel

@surface a

@column 1

```
1. {d}na3-na-'i-id _lugal_ babilax(|TIN.TIR|){ki} _nun_ na-a-du
2. re-e2-a-am za-ni-nu sza2 a-na t,e4-em _dingir-mesz_ pu-tuq-qu
3. e-em-qa2 mu-ut-ne2-en-nu-u2 mu-usz-te-ne2-'u-u2 asz-ra-a-ti3 _dingir-mesz gal-mesz_
4. et,-lu szu-us-su-mu bi-nu-ut _abgal dingir-mesz_ {d}marduk
5. na-ab-ni-it {d}e4-ru6 ba-na-a-ta gi-mir ma-al-ku
6. i-ti-it {d}mu-u2-a-ti a-pil e2-sag-il2 sa-niq mit-hur-tu2
7. bi-nu-tu {d}nin-szi-ku3 mu-du-u2 ba-nu-u2 ka-la-ma
8. ni-bi-it {d}nanna-ri!(NIR) be-lu a-gi-i mu-kal-li-im s,a-ad-du
9. sza u4-mi-szam-ma isz-te-ne2-e'-u2 pu-luh-tu4 _dingir-mesz gal-mesz_
10. a-na zi-in-na-a-ti e2-sag-il2 u3 e2-zi-da
11. ba-sza-a uz-na-a-szu
12. _dumu_ {disz}{d}na-bi-um-ba-lat-su-iq-bi _nun_ e-em-qa2 a-na-ku
13. _iri_ babil2{ki} a-na dam-qa-a-ti asz-te-ne2-e'-e
14. a-na e2-sag-il2 _e2-gal dingir-mesz gal-mesz_ szu-tu2-ra-ku _igi-sa2_-e
15. a-na e2-zi-da sza2-ad ba-la-t,u mim-ma szum-szu du-usz-sza2-ku
16. sza2 e2-mes-lam _e2_ qar-ra-du _dingir-mesz_ t,u2-uh-hu-da-ak hi-is,-bi
17. i3-nu-szu im-gur-{d}en-lil2 _{<<d>>en-lil2}> bad3_ babil2{ki}
18. isz-da-a-szu i-nu-szu-ma i-qu-pu i-ga-ru-szu
19. re-e-szi-szu it-ru-ur-ma ne2-me2-et-ta la i-szi
20. _bad3_ szu-a-ti <ana> du-un-nu-nim-ma ne2-me2-et-ta szu-ur2-szi-i
21. i-ga-ru-szu-gu-ur-pu-tin-ad-ka-e-ma
```

@column 2

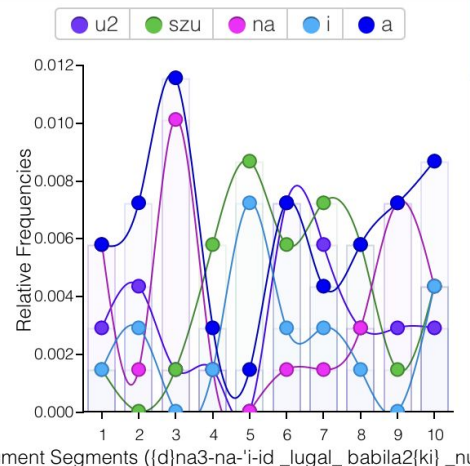
1. ma-ha-za!(A) re-esz-tu-u2 tuk-la-tu4# ba#-u2#-la-a!(ZA) tu2# i#-szi#-id#-[x]
2. u2-da-an-ni-in ki-ma kin-ne2-e u2-pat-tin-ma
3. sza-da-ni-isz u2-zaq-qi2-ir me2-la-a-szu
4. uhu2-mi-isz u2-szar-szi-id-su!(SZU) a-na tab-ra-a-tu2 asz-tak2-[x x]
5. szi-t,i-ir szu-mi sza2 _lugal_ mah-ri sza2 qe2-reb-szu ap-pal-[x]
6. it-ti szi-t,i-ir szu-mi-ia qer-ba-szu u2-ki-in a-na s, a#-[x x]
7. {d}en-lil2 _dingir-mesz_ {d}marduk sza2 qi2-bit-su ki-na-at#
8. be-lu ap-kal-lu4 _dingir-mesz_ szu-ur2-bu-u2 qu-[x x]
9. szi-ip-ri szu-a-ti ha-di-isz nap-lis-[x]
10. mi-im-mu-u2 e-te-ep-pu-szu li#-[x x]

{d}na3-na-'i-id _lugal_ babila2{ki} _nun_ na-a-du
re-e2-a-am za-ni-nu sza2 a-na t,e4-em _dingir-mesz_ pu-tuq-qu
e-em-qa2 mu-ut-ne2-en-nu-u2 mu-usz-te-ne2-'u-u2 asz-ra-a-ti3 _dingir-mesz gal-mesz_
et,-lu szu-us-su-mu bi-nu-ut _abgal dingir-mesz_ {d}marduk
na-ab-ni-it {d}e4-ru6 ba-na-a-ta gi-mir ma-al-ku
i-ti-it {d}mu-u2-a-ti a-pil e2-sag-il2 sa-niq mit-hur-tu2
bi-nu-tu {d}nin-szi-ku3 mu-du-u2 ba-nu-u2 ka-la-ma
ni-bi-it {d}nanna-ri be-lu a-gi-i mu-kal-li-im s,a-ad-du
sza u4-mi-szam-ma isz-te-ne2-e'-u2 pu-luh-tu4 _dingir-mesz gal-mesz_
a-na zi-in-na-a-ti e2-sag-il2 u3 e2-zi-da
ba-sza-a uz-na-a-szu
dumu {disz}{d}na-bi-um-ba-lat-su-iq-bi _nun_ e-em-qa2 a-na-ku
iri babila2{ki} a-na dam-qa-a-ti asz-te-ne2-e'-e
a-na e2-sag-il2 _e2-gal dingir-mesz gal-mesz_ szu-tu2-ra-ku _igi-sa2_-e
a-na e2-zi-da sza2-ad ba-la-t,u mim-ma szum-szu du-usz-sza2-ku
sza2 e2-mes-lam _e2_ qar-ra-du _dingir-mesz_ t,u2-uh-hu-da-ak hi-is,-bi
i3-nu-szu im-gur-{d}en-lil2 bad3_ babila2{ki}
isz-da-a-szu i-nu-szu-ma i-qu-pu i-ga-ru-szu
re-e-szi-szu it-ru-ur-ma ne2-me2-et-ta la i-szi
bad3 szu-a-ti <ana> du-un-nu-nim-ma ne2-me2-et-ta szu-ur2-szi-i



```
{
  {d}na3-na-'i-id _lugal_ babila2{ki}
  _nun_ na-a-...

  d)na3-na-'i-id _lugal_ babila2{ki} _nun_ na-a-
  du
  re-e2-a-am za-ni-nu sza2 a-na t,e4-em _dingir-
  mesz_ pu-tuq-qu
  e-em-qa2 mu-ut-ne2-en-nu-u2 mu-usz-te-ne2-
  'u-u2 asz-ra-a-ti3 _dingir-mesz gal-mesz_
  et,-lu szu-u-su-mu bi-nu-ut _abgal dingir-
  mesz_ {d}marduk
```



Terms:

This corpus has 1 document with 693 total words and 192 unique word forms.
Created about 5 minutes ago.

Vocabulary Density: 0.277

Readability Index: 24.001

Average Words Per Sentence: 693.0

Most frequent words in the corpus:

Items:

	Left	Right
<input type="checkbox"/>	{.. _lugal_ babila2{ki} _nun_ na-	a -du re-e2-a-am
<input type="checkbox"/>	{.. na-a-du re-e2-	a -am za-ni-nu sza2
<input type="checkbox"/>	{.. am za-ni-nu sza2	a -na t,e4-em _dingir
<input type="checkbox"/>	{.. ne2-'u-u2 asz-ra-	a -ti3 _dingir-mesz gal-mesz_
<input type="checkbox"/>	{.. d)e4-ru6 ba-na-	a -ta gi-mir ma-al
<input type="checkbox"/>	{.. ti-it {d}mu-u2-	a -ti a-pil e2-sac

expand

{ d } na3-na- ' i-id _lugal_ babila2 { ki } _nun_
na-a-du
re-e2-a-am za-ni-nu sza2 a-na t , e4-em _dingir-
mesz_ pu-tuq-qu
e-em-qa2 mu-ut-ne2-en-nu-u2 mu-usz-te-ne2- ' u-
u2 asz-ra-a-ti3 _dingir-mesz gal-mesz_
et , -lu szu-us-su-mu bi-nu-ut _abgal dingir-mesz_
{ d } marduk
na-ab-ni-it { d } e4-ru6 ba-na-a-ta gi-mir ma-al-ku
i-ti-it { d } mu-u2-a-ti a-pil e2-sag-il2 sa-niq mit-
hur-tu2
bi-nu-tu { d } nin-szi-ku3 mu-du-u2 ba-nu-u2 ka-la-
ma
ni-bi-it { d } nanna-ri be-lu a-gi-i mu-kal-li-im s , a-
ad-du
sza u4-mi-szam-ma isz-te-ne2-e ' -u2 pu-luh-tu4
dingir-mesz gal-mesz
a-na zi-in-na-a-ti e2-sag-il2 u3 e2-zi-da
ba-sza-a uz-na-a-szu
dumu { disz } { d } na-bi-um-ba-lat-su-iq-bi
nun e-em-qa2 a-na-ku

Nabonidus , king of Babylon , attentive prince ,
the shepherd who provides , the one who is
constantly attentive to the will of the gods , the
wise (and) pious one , the one who constantly
seeks out the shrines of the great gods , most
befitting warrior , creation of the sage of the gods
— the god Marduk — product of the goddess
Erua — creator of all rulers — selected by the
god Nabû — the heir of Esagil who controls (
cosmic) harmony — creation of the god Ninšiku
— the (all-) knowing creator of everything —
chosen by the god Nannāru — the lord of the
crown who makes astrological signs known — the
one who strives every day (to show) devotion to
the great gods (and) whose mind is focused on
provisioning Esagil and Ezida , son of Nabû-
balāssu-iqbi , wise prince , am I .

Text cleaning

- Removing “noise” from a text for analysis
 - Correcting errors, e.g. typographical errors, OCR artefacts
 - Removing metadata, e.g. line numbers or page numbers
- Significant part of any project working on digitising/digitised texts
- Clean data is task-specific

Text cleaning solutions

- By hand
- Online tools
- AI
- Regular expressions

What are Regular expressions?

What are regular expressions?

- Find-and-replace based on patterns not literal characters
- Eg validate an email address in an online form
- The 'find' part is useful for counting things
- The 'replace' part is useful for cleaning and preparing
- They don't do anything else!

Why learn regular expressions?

- They are (almost) ubiquitous in text editing software
- Any standard programming language has them
- They vary very little between implementations
- They are very modular: you can do a lot with a few chars
- The basic syntax is very minimal: not much to learn
- (But you will need to practise)

Regex syntax

Regular Expressions Cheat Sheet by [DaveChild](#)

A quick reference guide for regular expressions (regex), including symbols, ranges, grouping, assertions and some sample patterns to get you started.

Anchors

<code>^</code>	Start of string, or start of line in multi-line pattern
<code>\A</code>	Start of string
<code>\$</code>	End of string, or end of line in multi-line pattern
<code>\Z</code>	End of string
<code>\b</code>	Word boundary
<code>\B</code>	Not word boundary
<code>\<</code>	Start of word
<code>\></code>	End of word

Character Classes

`\c` Control character

Quantifiers

<code>*</code>	0 or more	<code>{3}</code>	Exactly 3
<code>+</code>	1 or more	<code>{3,}</code>	3 or more
<code>?</code>	0 or 1	<code>{3,5}</code>	3, 4 or 5

Add a `?` to a quantifier to make it ungreedy.

Escape Sequences






<code>\</code>	Escape following character
<code>\Q</code>	Begin literal sequence
<code>\E</code>	End literal sequence

"Escaping" is a way of treating characters which have a special meaning in regular expressions


Groups and Ranges


<code>.</code>	Any character except new line (<code>\n</code>)
<code>(a b)</code>	a or b
<code>(...)</code>	Group
<code>(?:...)</code>	Passive (non-capturing) group
<code>[abc]</code>	Range (a or b or c)
<code>[^abc]</code>	Not (a or b or c)
<code>[a-q]</code>	Lower case letter from a to q
<code>[A-Q]</code>	Upper case letter from A to Q
<code>[0-7]</code>	Digit from 0 to 7

<https://cheatography.com/davechild/cheat-sheets/regular-expressions/>





SAVE & SHARE


 Save new Regex **ctrl+s**


 Add to Community Libr...


FLAVOR


 **PCRE2 (PHP >=7.3)** ✓


 PCRE (PHP <7.3)


 ECMAScript (JavaScript)


 Python

 Golang


 Java 8


 .NET 7.0 (C#)


 Rust


 Regex Flavor Guide

FUNCTION


 **Match** ✓


 Substitution


 List


 Unit Tests

TOOLS

 Code Generator


 Regex Debugger

 Export Matches

 Benchmark Regex

REGULAR EXPRESSION

112 matches (826 steps, 19.78ms) ⓘ

`<.+>` / gm 

TEST STRING

```
<?xml*version="1.0"*encoding="UTF-8"?>
<?xml-model*href="http://epidoc.stoa.org/
schema/latest/tei-epidoc.rng"*
schematypens="http://relaxng.org/ns/
structure/1.0"?>
<?xml-model*href="http://epidoc.stoa.org/
schema/latest/tei-epidoc.rng"*
schematypens="http://purl.oclc.org/dsd/
schematron"?>
<TEI*xmlns="http://www.tei-c.org/ns/1.0"*
xml:space="preserve"*xml:lang="en">
...<teiHeader>
...<fileDesc>
...<titleStmt>
...<title>Philokales*first*
inscription</title>
...<author>
xml:id="tp">Theocharis*Petrrou</author>
...<author*xml:id="mx">Maria*
Xenaki</author>
...<editor*xml:id="mgp">*Maria*
G.*Parani</editor>
...</titleStmt>
...<publicationStmt>
```

EXPLANATION

▼ / `<.+>` / gm

< matches the character < with index 60₁₀ (3C₁₆ or 74₈) literally (case sensitive)

▼ . matches any character (except for line terminators) ⓘ

+ matches the previous token between one and unlimited times, as many times as possible, giving back as needed (greedy)

> matches the character > with index 62₁₀ (3E₁₆ or 76₈) literally (case sensitive)

▼ Global pattern flags

g modifier: global. All matches (don't return after first match)

m modifier: multi line. Causes ^ and \$ to match the begin/end of each line (not only begin/end of string)

MATCH INFORMATION

▼

Match 1 0-38

```
<?
xml*version="1
.0"*encoding="
UTF-8"?>
```

<https://regex101.com/>

Regex demos (live)

Python demo (live)

When not to use regex

Exercise




Understanding Regular Expressions

Doug Knox

In this lesson, we will use advanced find-and-replace capabilities in a word processing application in order to make use of structure in a brief historical document that is essentially a table in the form of prose.

 Peer-reviewed

 CC-BY 4.0

 Support PH

E D I T E D B Y

Adam Crymble

R E V I E W E D B Y

Dave Shepard

Patrick Burns

P U B L I S H E D | 2013-06-22

 <https://doi.org/10.46430/phen0033>

M O D I F I E D | 2020-05-12

D I F F I C U L T Y | Medium