CRMtex

Textual Entities Model

Definition of the CRMtex

An Extension of CIDOC CRM to Model Textual Entities

Proposal for approval by CIDOC CRM - SIG

Version 1.2
August 2022

Currently maintained by Francesca Murano and Achille Felicetti

Contributors:
Martin Doerr, Pavlos Fafalios, Achille Felicetti, Francesca Murano, Athanasios Velios

Index

1 Introduction

1.1 Scope

This document presents CRM*tex*, an extension of the CIDOC CRM created to support the study of ancient and handwritten documents, i.e., texts characterised by uniqueness since they have been produced without the use of techniques typical of the modern mechanized process of production. CRMtex aims to model information concerning ancient inscriptions (including coins, marks and stamps), papyri, medieval manuscripts, but also modern handwritten documents of any kind.
Furthermore, CRMtex proposes the use of the CIDOC CRM to encode this kind of documents and to model their scientific process of investigation to foster the integration with other cultural heritage research fields. After identifying the key concepts, assessing the available technologies, and analysing the entities provided by CIDOC CRM and its extensions, CRMtex introduces new classes and properties to address the needs of the disciplines involved (including epigraphy, papyrology, palaeography, and codicology).

Investigating written documentation

On present archaeological evidence, full writing appeared in Mesopotamia and Egypt around the end of the IV millennium BC. (cf. Feldherr and Hardy 2011). With the evolution of this technology, humans began to write texts on different supports using different techniques: inscriptions, papyri, manuscripts, and other similar documents.
Although from the semiotic point of view (see below) the mechanism of production of written texts follows a unique approach (regardless the used supports, techniques, etc), traditionally, the study of ancient texts falls within different disciplines, generally grown around the specific characteristics of each class of documents (e.g., papyrology for the study of papyri, epigraphy for inscriptions and palaeography for the study of ancient manuscripts). Nevertheless, an interdisciplinary approach is essential, and the identification of common elements is paramount to confer uniformity and interoperability to all these disciplines, as well as to exploit complementary skills from different approaches.

What should be observed, specifically in texts for which this model was designed, is the relationship between the text and its support. In comparison to modern printed or digital texts, this kind of texts is typically characterised by its uniqueness, being the result of manual work rather than a mechanised process, as occurs since the invention of modern printing techniques.
Such characteristics make the study and digitisation of this type of documentation particularly arduous: the close relationship between the text and its support requires careful analysis since they are inextricably linked to form a unique object of study.
In the ancient world, nevertheless, some types of inscriptions were created through mechanised processes, such as the legends of coins, medals, stamps, and seals. The uniqueness of the written text remains unchanged in this case also, since it is characterised by the peculiar history of the support, which in most of the cases is a cultural object having significance also for other disciplines (e.g., numismatics, archaeology, etc.).

The first aim of this extension is therefore to identify and define in a clear and unambiguous way the main entities involved in the study and edition of ancient and other

handwritten texts and then to describe them by means of appropriate ontological instruments in a multidisciplinary perspective.

In addition to dealing with text as an object, our model also focuses on the aspects of the research and provides classes and relationships to describe the typical operations that scholars from different disciplines put in place to gain knowledge about texts. It is evident, in this perspective, that the study of ancient texts typically starts from the analysis of the physical characteristics of the individual text itself before moving to the investigation of their archaeological, palaeographic, linguistic, and historical features (see Figure 2).

What is said and what is written

"Writing is one of the most significant cultural accomplishments of human beings" (cf. Rogers 2005), since it offers humans new semiotic resources, allowing to communicate either at a distant place and time, and, thus, to hand down the memory, written texts being more enduring than spoken utterances.
Although every speech can be transposed into an equivalent written message, and vice versa, according to a common formulation of the relationship between speech and writing, the spoken language is prior to writing (cf. Lyons 1972), in the sense that writing results from the transference of the language from a primary phonic medium to a secondary graphic medium (cf. Lyons 1977: 65).
Writing originated as a representation of speech, "as the use of graphic marks to represent specific linguistic utterances" (Rogers 2005). According to Ferdinand de Saussure (1983), "a language and its written form constitute two separate systems of signs. The sole reason for the existence of the latter is to represent the former".
Although writing can be examined from a variety of perspectives, being it applicable to different human activities (Harris 1995), the theory on which CRMtex is based for the analysis of writing is that of semiotics. In a semiotic perspective, language and writing are codes (i.e., systems of signs) and the transmission of a message is an encoding/decoding process: the formation of a message by the sender is an encoding and the interpretation of the message is a decoding by the receiver. Coding consists in assigning the appropriate expression to a given content; the decoding in identifying the content starting from the expression. In this theorical framework (and desiring to simplify a very complex matter), writing is a secondary code, having as its content the expression of another code (i.e., the language).

Writing, therefore, appears as a code requiring an encoding process by the creator or writer and a decoding one by the receiver or reader to be properly understood (see Figure 1). It is worth considering that in writing (characteristic in common with speech) every component (sign) possesses a dual nature, one physical and another conceptual: regardless techniques and types of supports, writing "involves the physical production of variable tokens representing invariant types" (i.e., the ideal shapes of the signs of a writing system) (Coulmas 1999: 193). Thus, for the analysis of written texts it is necessary to distinguish the concrete, physical, individual realization performed by a single person on a specific occasion (e.g., the specific unique sequence of marks I wrote on a paper with my pen to take down a note), and the abstract level concerning the mental

knowledge and rules pertinent to a particular writing system, allowing the process of recognizability of the material mark with an "ideal" sign on the basis of a sameness principle. In brief the semiotic process underling the writing allows the identification of my personal "A" mark, independently from the peculiar shape I give to it, as the LATIN CAPITAL LETTER A.

Glyphs and Graphemes

The physical elements (glyphs or graphs) composing a written text constitute the material manifestations of the graphemes, i.e., the abstract entities of a writing system (cf. Coulmas 1999). According to the common definitions, a graph is the minimal formal unit of written language on the level of handwriting or print and a grapheme is the minimal functional distinctive unit of writing on whatever structural level of language the writing system operates (Coulmas 1999; Pulgram 1976).
For a typological study of writing systems, scholars recognize a broad distinction between glottographic and non-glottographic (i.e. pure semasiographic) writing systems, "depending on whether the formation and interpretation of texts presupposes knowledge of a particular language" (Harris 1995: 95).
Concerning the glottographic systems (that non-glottographic writing systems in the narrower sense exist is very disputed), the "theory commonly adopted by linguists distinguishes different kinds of writing system according to which units in the spoken language appear to have been selected as the basic units for representation in writing" (Harris 1995: 95). According to Pulgram (1976: 2-3) a grapheme represents the minimal unit of some level: "in reducing a language to writing, that is, in making visible marks that evoke or recall linguistic performance, it would seem that each mark must represent a syntagmeme or a lexeme or a morpheme or a phoneme or whatever other kind of unit the inventor of the system may chose as his basis".
In glottographic systems scholars recognize a difference "between logographic scripts, which assign distinct marks to meaningful units of a language, i.e., words or morphemes, and phonographic scripts which represent phonological units of one size or another" (Sampson 2016; cf. Sampson 1985 and Rogers 2005).
To better clarify: in principle, in an alphabetic writing system, e.g., the Latin alphabet, including the consonantal ones (i.e., the abjads as the Arabic alphabet), the basic unit of representation is the phoneme. Both in a syllabic and in an alphasyllabic writing system (i.e., respectively in syllabary as the Mycenean or the Japanese systems, and abugida, as the Sanskrit or the Thai systems) basic unit is the syllable. In a logographic writing system, as (part of) the Egyptian hieroglyphic or the modern Chinese system the basic unit of representation is a grammatical/lexical unit (i.e., a morpheme or a word) (cf. Daniels and Bright 1996; Borgwaldt and Joyce 2013). We propose some examples. In a Latin inscription, each mark inscribed on the stone (i.e., each glyphs) represents a corresponding grapheme in the Latin writing system (which in turn stands for a phoneme): e.g., the first five glyphs of the last line in fig. 4 represent the graphemes <a>, <r>, <c>, <u> and <m> of the Latin alphabet, and in turn these graphemes codify the following sounds of the Latin language (phonemes): /a/, /r/, /k/, /u/ and /m/ (Lat. *arcum* 'arch' acc. sing.). In Mycenaean Linear B and in Old Persian cuneiform inscriptions, glyphs represent (for the most part) syllabograms, i.e., the graphemes representing a syllable, not a single sound. E.g., the first sequence visible on the inscription

from the Palace of Darius the Great in Persepolis (cf. here) represents the seven graphemes of the Old Persian writing system 𐎠𐎡𐎢𐎣𐎤𐎥𐎦 corresponding to the seven syllables /da/, /a/, /ra/, /ja/, /va/, /u/ and /ʃa/. In an Egyptian hieroglyphic text, glyphs may represent syllabic, alphabetic or ideographic elements, i.e., the elements standing for lexical/semantic units.

Therefore, writing systems show over the time deviations from the ideal 1:1 correspondence between units of the language (whatever they are) and units of the writing system (grapheme), determining spelling conventions as product of changes to which linguistic systems are subjected in diachrony.
This phenomenon is particularly evident in phonographic systems, because of the diachronic phonetical variations. From this, it follows that, for example, in English, many discrepancies appears between spelling and phonetic values: e.g., the grapheme <i> stands for various phonemes: /ɪ/ (as in *him*), /ʌɪ/ (as in *time*), /i/ (as in *police*), /a/ (as in *timbre*); vice versa, the phoneme /f/ can be represented with <f> (as in *film*), <ph> (as in *philology*) or <gh> (as in *enough*).
For scientific purposes, the International Phonetic Alphabet has been devised as a standardized representation of speech sounds in written form, having a 1:1 correspondence between phonological unit and IPA symbols.

Recognising, reading and understanding the text

Reading refers to the semiotic procedure of decoding a written text, and therefore of deriving meaning from it (i.e., understanding it). Reading is "a highly complex activity involving the interplay of visual-perceptual, linguistic and conceptual systems" (Coulmas 1999: 430).
From a semiotic point of view, according to the communication theories, a complete retrieval of the information (i.e., reading of the written message) presupposes the code sharing by sender and receiver (Jakobson 1960). In history, the code that links writing to meaning or sound or both may have been lost and scholars, for scientific purposes, must recover it/them.

Scholars propose various models of the reading process, based on the identification of "the perceptual and cognitive stages and activities leading from visual input to understanding the content of the written message" (Coulmas 1999: 432), and distinguish some stages, from the eye fixation to the character identification, to the word recognition, to the association of meanings and the application of linguistic rules, finally to the application of phonological rules and the assignment of a phonetic form.

The reading process can be carried out for scientific purposes, to analyse and study the text according to different disciplinary perspectives. Although all writing is made to be read, their reading/comprehension depends on the degree of the initial knowledge of the reader (in the case of CRMtex the scientific community).
In the case of languages and writing systems that are no longer in use, in fact, it is possible that scholars are unable to entirely decode the elements, i.e., to establish the value that those elements have within the system. A case of this kind is constituted by the Linear A and the writing of the Phaistos disc, of which the linguistic systems they represent are unknown.

According to the aim of the model, regardless the cases in which the observation of visual items on a surface does not determine the recognition of a text and concerning only the cases of the observation of a text, we consider the following stages of the decoding process:
1. character identification: the process of identifying visual items as elements of a writing system; it is a necessary although not a sufficient condition of reading. Decoding processes that stop at this stage in the scientific field are due to the lack of knowledge of the used language (there is no code sharing between sender/writer and receiver/reader). An example is the current state of knowledge of the writing of the Phaistos disc;
2. signs recognition: the process of identifying element of a writing system known to the reader. At this stage the reader knows or can reconstruct the pronunciation and recognize the words, but the knowledge of the language is insufficient to have a complete linguistic comprehension of the text.
The deciphering of the signs can be achieved if the linguistic system represented is known; this is the case of Linear B, whose deciphering came after the understanding it represents a Greek language.
Since the writing systems have genealogical relationships with other known systems, it is possible that the writing systems do not present deciphering problems (so the scholar is able to attribute a rough value to the signs), even when the linguistic system it is not yet known. This is the case of the Etruscan writing system, which was deciphered from the origins of Etruscology, the Etruscan alphabet deriving from the Euboean one, although knowledge of the language (i.e., the understanding of the texts) is the result of a long study process that still presents uncertainties;
3. reading properly said: the process of associating the text with a complete linguistic meaning (cf. Coulmas 1999: 432).
On the level of the linguistic sounds, it will be the decoders (readers, including scholars), who from time to time, on the basis of the knowledge of the linguistic system, will attribute to each sign or group of signs the adequate (or reconstructed) phonetic value, also on the basis of spelling conventions in place in a given graphic system at a given historical moment, since the spelling rules can change over time, even if less quickly than the linguistic system does.

For the purposes of modelling the textual entity within the various disciplines for which CRMtex has been designed, within the model we distinguish two classes of text decoding depending on whether it is a proper reading or not.

For the goals of the study of texts, the reading activity requires a scientific autoptic examination of the text as preparatory action for the study. An autoptic examination consists of an accurate analysis of the surface and the signs and prescribes the use of specific tools and procedures, for establishing as faithfully as possible the exact value of each sign drawn or applied on the physical feature.

Reproductions, transcriptions and transliterations of a text

For research and scientific dissemination purposes, it is possible that there is a need to have a reproduction of text, also transposing it according to a writing system different from the original one.
According to their scientific purposes, scholars distinguish various stages:

1. an exact reproduction of the visual items recorded on a text (fac-simile). An example is the drawing of the inscription of Darius the Great in Persepolis published here;
2. a reproduction of the recognised graphemes of a text using the same writing system (transcription in a broader sense). An example is the text of the Dreros Law from Crete published by the AXON project (here)
3. a conversion (i.e., re-encoding) of the recognised graphemes of a text using a different writing system according to a 1:1 (i.e., unique and unambiguous) conversion (transliteration). Because of this 1:1 conversion this operation is reversible, allowing an "automatic" and unambiguous recreation of the original.
Since the purpose of transliteration is to enable those not familiar with a writing system in which a text is encoded to read it, commonly the Latin alphabet is used. An example is the text in Latin alphabet of the Ancient South Arabian inscription as-Sawdāʾ 49 published by the DASI project (here).
Transliteration conventions for writing systems structurally identical (e.g., alphabets), do not pose difficulties; in turn, conventions for writing systems of different type are more problematic.
In case of texts written in a non-alphabetic system, the conversion in Latin alphabet can involve linguistic elements broader than a phoneme, notwithstanding the 1:1 relation between the graphemes of source writing system and the Latin encoding. An example is the transliteration <da-a-ra-ya-va-u-ša> of the first sequence of inscription of Darius the Great in Persepolis published here, where a grapheme of the Persian syllabary corresponds to a syllable univocally and conventionally referred to in Latin script (e.g., <𐎭> → <da>).
For scientific purposes competing systems are in use in different disciplines, but each transliteration is consistent for a specific study field (e.g., Biblicists and linguists use different systems for transliterating Hebrew in Latin alphabet). Standards, as the ISO and BGN/PCGN, define the transliteration rules and are widely used to overcome these divergences.
4. a re-encoding of the recognised graphemes of a text using a different writing system according to a phonological (and even spelling) criterion (transcription in a narrower sense). For example, the name of Euboea region is Εύβοια in Greek alphabet; the sequence has transliteration 'Euboia' in Latin script but has transcription 'Evia', according to modern Greek pronunciation; notice that transcription is based on the phonetics, thus pronunciation problems can arise: for example an English speaker might read 'Evia' as [ˈɪvɪa] instead of [ˈɛvia], and possibly transcribe 'Ivia'). Another example: the name of the Russian composer Чайко́вский is transliterated according to the modern transliterations of Russian ISO 9 standard Čajkovskij; in turn the name is anglicized (i.e., transcribed according to the English system) as *Tchaikovsky* or *Chajkovskij*, etc., while in German is more common the transcription *Tschaikowskyi* and in French *Tschaïkowsky*.
For scientific purposes a re-encoding of this type is useful in case of text written in a non-alphabetic system, especially when, in composing words, the elements of the writing system do not match entirely with the actual phonetic structure of the represented word. An example is the transcription of the inscription of Darius the Great in Persepolis published here using the Latin alphabet, where each word is re-encoded taking into account the actual pronunciation regardless of how it is written in the original text (e.g. the first sequence reported above is transcribed Dârayavauš) .
A particular case is the conversion according to the International Phonetic Alphabet (IPA). Even though it consists of a change of the writing system employed, the use of

the IPA has as specific purpose the reproduction of the exact pronunciation of the words.

Written text segments

Scholars of different disciplines, on the basis of the requirements of their study, need to identify and focus their attention on different types of text segments, in order to describe their physical conditions (form, layout, etc.), to verify their legibility and particular phenomena (e.g., linguistic or palaeographic) connected to them, etc. For this reason, in designing the entities of CRMtex we created the class TX7 allowing the investigation of the interconnections existing between the text and its parts. Examples of text segments are columns, sections, paragraphs, but also single words or letters, or other specific components of the written text that scholars need for their purposes.
In this way it is possible to assign specific issues to the individual segments, independently of the text in its entirety. In fact, particular production (i.e. TX2 *Writing*) or destruction (E6) events can be associated with single segments, as in the case of letters or words damaged or worn out due to deterioration or human interventions.

Specifications about conditions (E3) for documenting the state of each textual part during the observation process (S4) can be easily stated as well. This allows scholars to document different events for the investigated parts in a more precise way and to assign observations and interpretations to them (see Fig. 3).

Style and other palaeographic features

Since the stylistic variations of hand-written texts are constitutive (e.g., an 'A' can appear as uppercase, lowercase, italics, round, printed or written by hand, or in different font families), a palaeographic study of stylistic variations has great importance in the description of written texts, using different styles for different purposes or at different times and places.

This approach is fundamental for the determination of the dating and provenance of the texts, especially in reference to the styles developed in certain centres (for example, in the scriptoria of the monasteries). It is also relevant for the description of all the entities of a given epoch and place, e.g., the Ptolemaic cursive of the Hellenistic Egypt, the capital uncial script (3rd-8th cent. AD), used both for Greek and Latin alphabets, or the more recent Carolingian minuscule, used from the beginning of the 8th cent. AD.
Therefore, in palaeography the concepts of stylistic class, style and canon are paramount to underline different meaningful observable aspects. The specific study of these stylistic variations needs to be properly addressed.

Palaeography uses different concepts, including aspects of the style, writing direction and other features related to the physical way the text is written and arranged.

1.2 Status

CRMtex is the result of collaboration between scholars of many cultural heritage institutions. The first need that the model attempts to meet is to create a common ground for the integration and interoperability of records concerning ancient texts on every level, from the description of the supports and carried texts to the management of the documentation produced by various institutions using national and institutional standards (e.g., TEI/EpiDoc). This document describes a community model, under approval by CRM SIG as being formally and methodologically compatible with CIDOC CRM. However, in a broader sense, it is always open to any possible integration and addition that may become necessary as a result of its practical use on real problems on a large scale. The model is intended to be maintained and promoted as an international standard.

1.3 Naming Convention

All the classes declared were given both a name and an identifier constructed according to the conventions used in the CIDOC CRM model. For classes that identifier consists of the letter TX followed by a number. Resulting properties were also given a name and an identifier, constructed according to the same conventions. That identifier consists of the letters TXP followed by a number, which in turn is followed by the letter "i" every time the property is mentioned "backwards", i.e., from target to domain (inverse link). "TX" and "TXP" do not have any other meaning. They correspond respectively to letters "E" and "P" in the CIDOC CRM naming conventions, where "E" originally meant "entity" (although the CIDOC CRM "entities" are now consistently called "classes"), and "P" means "property". Whenever CIDOC CRM classes are used in our model, they are named by the name they have in the original CIDOC CRM. CRMsci classes and properties are referred with their respective names, classes denoted by S and properties by O.

2 Classes and properties hierarchies

The CIDOC CRM model declares no “attributes” at all (except implicitly in its “scope notes” for classes), but regards any information element as a “property” (or “relationship”) between two classes. The semantics are therefore rendered as properties, according to the same principles as the CIDOC CRM model.

Although they do not provide comprehensive definitions, compact mono hierarchical presentations of the class and property IsA hierarchies have been found to significantly aid in the comprehension and navigation of the model and are therefore provided below.

The class hierarchy presented below has the following format:

- Each line begins with a unique class identifier, consisting of a number preceded by the appropriate letter “E”, “TX”, “S”

- A series of hyphens (“-”) follows the unique class identifier, indicating the hierarchical position of the class in the IsA hierarchy.

- The English name of the class appears to the right of the hyphens.

- The index is ordered by hierarchical level, in a “depth first” manner, from the smaller to the larger sub hierarchies.

- Classes that appear in more than one position in the class hierarchy as a result of multiple inheritance are shown in an italic typeface.

2.1 CRMtex class hierarchy, aligned with portions from the CRMsci, LRMoo and the CIDOC CRM class hierarchies

This class hierarchy lists:
• all classes declared in Ancient Text model (CRMtex)
• all classes declared in CRM*sci* and CIDOC CRM that are declared as superclasses of classes declared in the Ancient Text Model,
• all classes declared in CRM*sci* or CIDOC CRM that are either domain or range for a property declared in the Ancient Text Model,
• all classes declared in CRM*sci* and CIDOC CRM that are either domain or range for a property declared in Ancient Text Model or CIDOC CRM that is declared as super-property of a property declared in the Ancient Text Model,
• all classes declared in CRM*sci* and CIDOC CRM that are either domain or range for a property that is part of a complete path of which a property declared in Ancient Text Model is declared to be a shortcut.

E1 CRM Entity
S15 - Observable Entity
E2 - - Temporal Entity
E5 - - - Event
E7 - - - - Activity
TX6- - - - - - Transcription
E13 - - - - - Attribute Assignment
S4 - - - - - - Observation
TX5- - - - - - - Reading
E63 - - - - Beginning Of Existence
E12 - - - - - Production
F28 - - - - - - Expression Creation
TX2- - - - - - - Writing
E77 - - Persistent Item
E70 - - - Thing
E72 - - - - Legal Object
E18 - - - - - Physical Thing
E26 - - - - - - Physical Feature
E25 - - - - - - - Man-made Feature
TX1 - - - - - - - - Written Text
TX7- - - - - - - - - Written Text Segment
TX9 - - - - - - - - Glyph
TX4- - - - - - - - Writing Field
E71 - - - - Man-made Thing
E28 - - - - - Conceptual Object
E90 - - - - - - Symbolic Object

TX8 - - - - - - - Grapheme
E73 - - - - - - - Information Object
E29 - - - - - - - - Design or Procedure
TX3 - - - - - - - - - Writing System
TX10 - - - - - - - - - Style

2.2 CRMtex property hierarchy, aligned with portions from the CRMsci and the CIDOC CRM property hierarchies

This property hierarchy lists:
• all properties declared in Ancient Text Model,
• all properties declared in CRM*sci* and CIDOC CRM that are declared as superproperties of properties declared in Ancient Text Model,
• all properties declared in CRM*sci* and CIDOC CRM that are part of a complete path of which a property declared in Ancient Text Model, is declared to be a shortcut.

Property id	Property Name	Entity - Domain	Entity - Range
TXP1	used writing system (writing system used for)	TX2 Writing	TX3 Writing System
TXP2	includes (is included within)	TX4 Writing Field	TX1 Written Text
TXP3	rendered (is rendered by)	TX6 Transcription	TX5 Reading
TXP4	has segment (is segment of)	TX1 Written Text	TX7 Written Text Segment
TXP5	wrote (was written by)	TX2 Writing	TX1 Written Text
TXP6	encodes (is encoding of)	TX3 Writing System	E33 Linguistic Object
TXP7	has item (is item of)	TX3 Writing System	TX8 Grapheme
TXP8	has component (is component of)	TX1 Written Text	TX9 Glyph
TXP9	is encoded using (was used to encode)	TX1 Written Text	TX3 Writing System
TXP10	read (was read by)	TX5 Reading	TX1 Written Text
TXP11	transcribed (was transcribed by)	TX6 Transcription	TX8 Grapheme
TXP12	has style (is style of)	TX1 Written Text	TX10 Style

3 Graphical overview

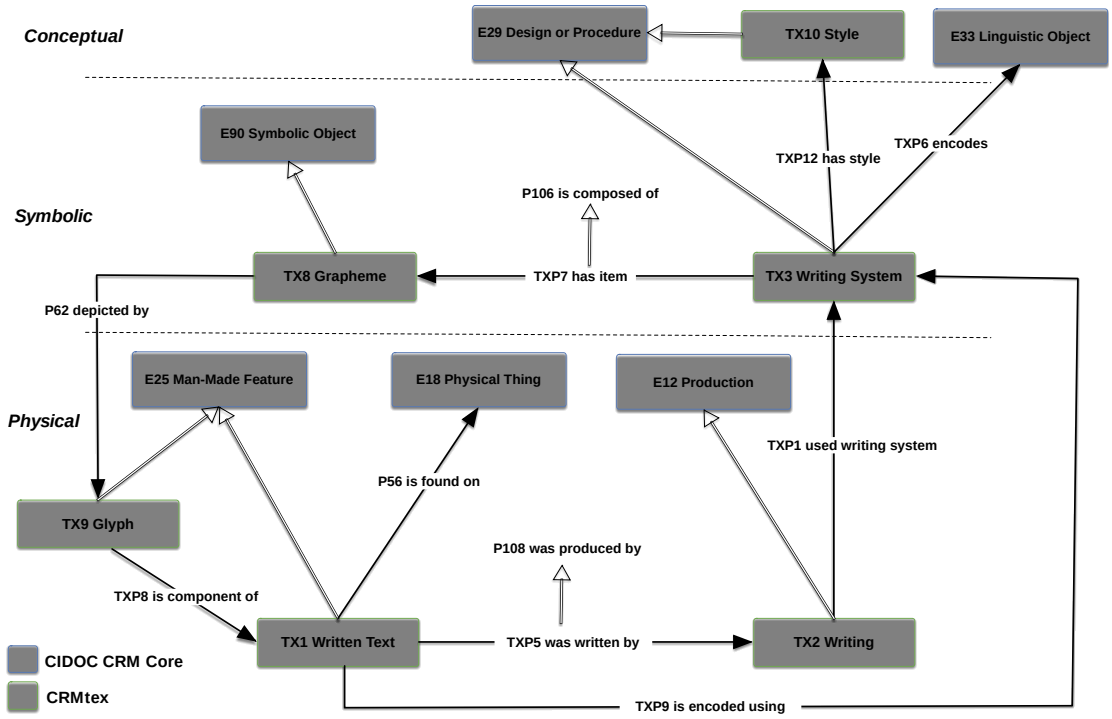

Figure 1: Text entities and text production in CRMtex

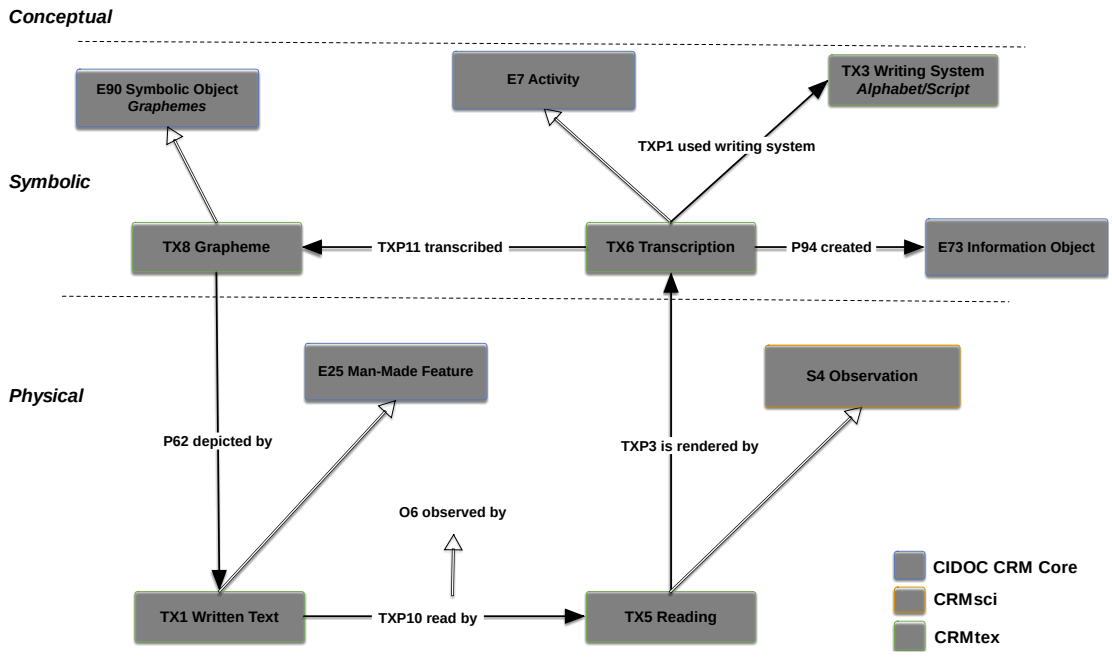

Figure 2: The process of investigation of ancient texts in CRMtex

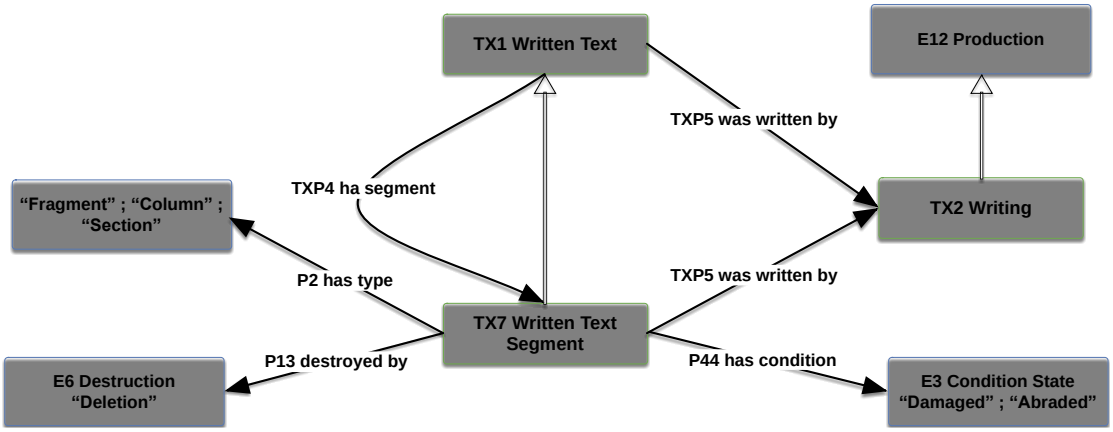

Figure 3: Written Text and Written Text Segments in CRMtex.

4 Classes and properties usage examples

The following example is intended to illustrate how CRMtex classes and properties could be used to encode, for instance, epigraphic information. The inscriptions on the Arch of Constantine, one of the most famous ancient monuments in Rome, have been chosen as examples of an ancient text occurring on a physical carrier in order to show how they can be semantically described in relation to the archaeological object carrying them.

The monument, still located in its original position between the Colosseum and the Roman Forum, is a triumphal marble arch – the largest monument of this kind in Roman Empire – dedicated in 315/316 A.D. by the Roman Senate to the emperor Constantine after his victory over Maxentius in the Battle of the Milvian Bridge in 312 A.D.

Among the other decorations (including statues, panels, reliefs and similar decorative material), the arch carries, on its attic, two identical inscriptions (reference number: CIL VI 1139), originally inlaid with gilded bronze letters, explaining the reason of its construction.
As of today, the bronze letters are lost and only the large cuttings in the marble, in which the bronze letters sat, remain. The inscription (Figure 5) is repeated, identically, on the South and North faces of the arch's attic. A transcription and a translation in English of the same inscription is presented below.

Figure 4: The inscription on the South face on the attic of the Arch of Constantine.

Inscription Transcription

IMP(ERATORI) · CAES(ARI) · FL(AVIO) · CONSTANTINO · MAXIMO · P(IO) · F(ELICI) · AVGUSTO · S(ENATUS) · P(OPULUS) · Q(UE) · R(OMANUS) · QVOD · INSTINCTV · DIVINITATIS · MENTIS · MAGNITVDINE · CVM · EXERCITV · SVO · TAM · DE · TYRANNO · QVAM · DE · OMNI · EIVS · FACTIONE · VNO · TEMPORE · IVSTIS · REMPVBLICAM · VLTVS · EST · ARMIS · ARCVM · TRIVMPHIS · INSIGNEM · DICAVIT

Inscription Translation

To the Emperor Caesar Flavius Constantine, the Greatest, Pius, Felix, Augustus: inspired by (a) divinity, in the greatness of his mind, he used his army to save the state by the just force of arms from a tyrant on the one hand and every kind of factionalism on

the other; therefore, the Senate and the People of Rome have dedicated this exceptional arch to his triumphs.

CRMtex description of the text

The Arch is an archaeological object and according to the CIDOC CRM it can be represented as an instance of the E22 *Man-made Object* class. The monument, made of marble, was overall intended to commemorate the emperor and not to carry the various the inscriptions present on it. A writing event (TX2) can be assigned to the inscriptions, thus it is always possible to distinguish the production event of the monument from that one of the inscriptions when it is needed.

CRMtex can be used to describe the two inscriptions appearing on the arch and relate them to the monument via the P56 *bears feature (is found on)* property. Each of the two inscriptions can be rendered as a TX1 *Written Text*, being the physical features intended to carry a particular significance. A TX2 *Writing event* can be specified for each TX1 via the TXP5 *was written by* property to render the production of the cuttings made to host the bronze letters. Since there are two inscriptions, we have the opportunity, this way, to distinguish the two processes that led to the production of each of them.

A TX4 *Writing Field* class can be used to describe the portion of the surface of the arch reserved by the builders and appositely arranged for accommodating the inscription, in order to highlight it from the other parts of the object and to enhance its readability. Thus, the CRMtex encoding in this case will include two TX4s instances.

The linguistic message to be conveyed (E33 Linguistic Object) encoded by means of a language (E56 *Language*) and by means of the writing system (TX3 *Writing System*) this language uses. From this follows that the TX1 *Written Text* class is the concrete graphical manifestation (i.e. a set of signs – in this case the engraved letters – we can read on the stone) of the conceptual level of encoding a linguistic expression through the semiotic activity of writing (TX2 *Writing*) by means of a TX3 *Writing System* (in this case, Latin alphabet) and of the graphemes (TX8) composing it.

The reading of a text, from a semiotic point of view, is a decoding activity. In CRMtex a reading – specially carried out for scientific purposes – can be documented using the TX5 *Reading* class underlying the scientific nature of the investigation.

In fact, over the centuries, the arch of Constantine has been investigated thousands of times by scholars from all over the world and also reproduced by famous illustrators such as Giovan Battista Piranesi. Also, the inscriptions have been studied and transcribed several times in order to understand its nature, clarify the meaning of each section and improve its historical comprehension so as to put it in direct relation with the events that determined its creation. For this type of activity, specific classes and properties. The transcription of the text(s) present in *Corpus Inscriptionum Latinarum* (CIL VI 1139), for instance, can be represented via the TX6 *Transcription* class, while the analysis of the same inscription(s) carried out by Rodolfo Lanciani in 1892 [6] can be documented using the reading (TX5) class. Reading and transcription (TX6) activities can be related via the TXP3 *rendered* property, inherited by CIDOC CRM core.

The TX7 *Written Text Segment* class can be used to highlight specific portions of text on which the study focuses, on which specific phenomena appear or from which it is possible to derive special meanings. Figure 5 shows a CRMtex conceptualisation of the South inscriptions on the Arch of Constantine.

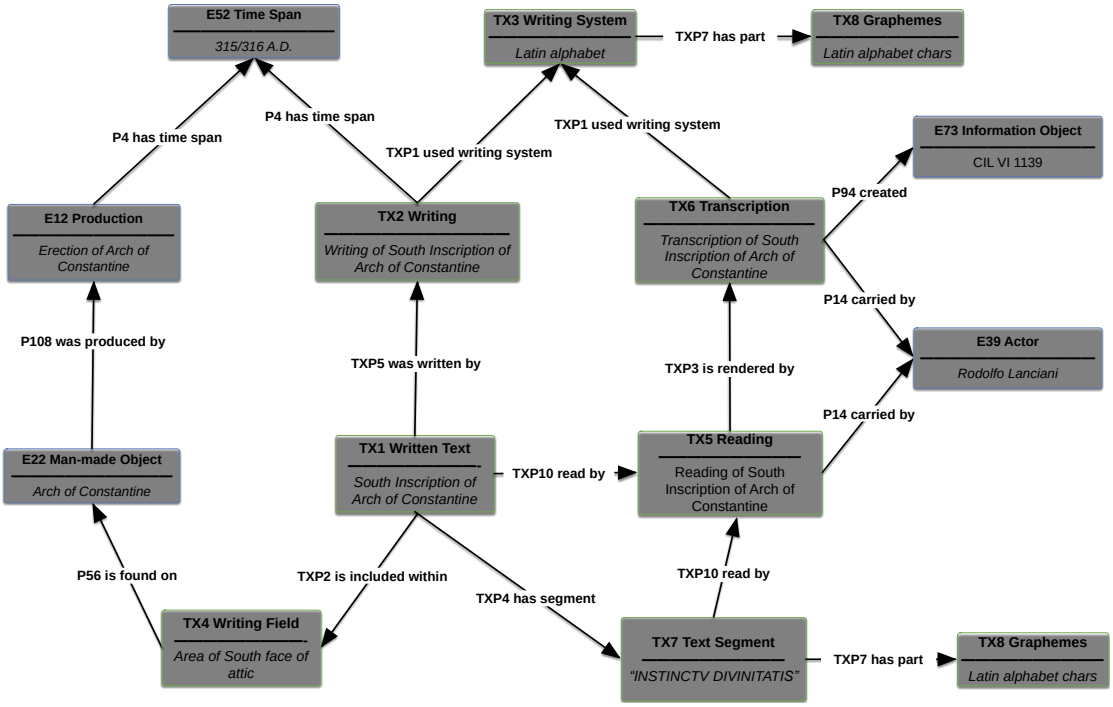

Figure 5: CRMtex encoding of one of the inscriptions (South) on the Arch of Constantine.

5 CRMtex - Classes and properties

5.1 CRMtex - Classes declarations

The classes are comprehensively declared in this section using the following format:

- Class names are presented as headings in bold face, preceded by the class's unique identifier;
- The line "Subclass of:" declares the superclass of the class from which it inherits properties;
- The line "Superclass of:" is a cross-reference to the subclasses of this class;
- The line "Scope note:" contains the textual definition of the concept the class represents;
- The line "Examples:" contains a bulleted list of examples of instances of this class.
- The line "Properties:" declares the list of the class's properties;
- Each property is represented by its unique identifier, its forward name, and the range class that it links to, separated by colons;
- Inherited properties are not represented;
- Properties of properties, if they exist, are provided indented and in parentheses beneath their respective domain property.

TX1 Written Text

Subclass of: E25 Man-Made Feature

Scope Note: This class comprises visible or tactile marks (called glyphs or graphs), which relate in a systematic way to units of speech, intentionally traced (i.e., "written") on some kind of physical support by using specific techniques and tools, with the purpose of conveying a message towards a given receiver or group of receivers.

Examples:

- The signs composing the inscription engraved on the South side of the attic of the Arch of Constantine (E22) in Rome (see section 1.3.1).
- The signs composing the text written on papyrus PSI XIII 1304 containing the so-called *Hellenica Oxyrhynchia* (TM 59482: https://trismegistos.org/text/59482).
-

In First Order Logic:

TX1(x) ⊃ E25(x)

Properties:

TXP4 has segment (is segment of): TX7 Written Text Segment
TXP8 has component (is component of): TX9 Glyph
TXP9 is encoded using (was used to encode): TX3 Writing System

TX2 Writing

Subclass of: F28 Expression Creation
Superclass of:
Scope Note: This class describes the activity of communicate information by means of permanent, visible marks in a non-mechanical way, using various techniques (painting, sculpture, etc.) and by means of specific tools, on a given support..

Examples:

- The process of engraving in the marble of the inscription (TX1) placed on the South attic of the Arch of Constantine (E22) in Rome (see section 1.3.1).

In First Order Logic:

TX2(x) ⊃ F28(x)

Properties:

TXP1 used writing system (writing system used by): TX3 Writing System

TXP5 wrote (was written by): TX1 Written Text

TX3 Writing System

Subclass of: E29 Design or Procedure
Superclass of:
Scope Note: This class represents conventional, symbolic system consisting of set of visible or tactile signs (graphemes, TX8) designed to represent units of a natural language with the purpose of recording and transmitting information. A complete retrieval of the transmitted messages requires a shared knowledge, between writers and readers, of the encoded language, the writing system elements and its encoding rules.
It is used to produce a TX1 Written Text during a TX2 Writing event.

Examples:

- The Latin alphabet used to encode the signs (TX1) composing the text (E33) of the inscription in Latin language occurring on the Arch of Constantine (E22).
- The Cypriot syllabary used in Iron Age Cyprus for codifying the Arcado-Cypriot dialect.
- The Chinese (Han) script used by Wang Xizhi to write the manuscript *Lanting Xu* ("Orchid Pavilion Preface").

In First Order Logic:

TX3(x) ⊃ E29(x)

Properties:

TXP6 encodes (is encoding of): E33 Linguistic Object
TXP7 has item (is item of): TX8 Grapheme

TX4 Writing Field

Subclass of: E25 Man-Made Feature
Superclass of:
Scope Note: This class describes the portion of the physical carrier arranged and usually reserved and delimited for the purpose of accommodating a written text, highlighting and isolating it from the other parts of the object to which it belongs, enhancing and guaranteeing its readability. This entity is paramount specially in epigraphy, in which a specific element called "epigraphic field" has been defined by the discipline itself. Its importance is also evident in papyrology and codicology, where a clear distinction between area(s) containing the written text and empty parts

of the support (margins, *intercolumnia*, etc.) is significant for the definition of styles and periods of the document.

Examples:

- The portion of the marble tombstone (E22) of M. Helvius Geminus from Ephesus reserved for accommodating the inscription (TX1).

In First Order Logic:

TX4(x) ⊃ E25(x)

Properties:

TXP2 includes (is included within): TX1 Written Text

TXxxx Reading

Subclass of: I1 Argumentation
Superclass of:
Scope Note: This class describes the complete intellectual activity, involving the interaction of visual-perceptual, linguistic, and conceptual systems, leading from text recognition (TX5) until its association with a complete linguistic meaning.

Examples:

- The reading of the South inscription (TX1) on the Arch of Constantine (E22) made by Rodolfo Lanciani between 1893 and 1901.
- The reading of the Greek text present on the Derveni papyrus (E22).

In First Order Logic:

TX5(x) ⊃ S4(x)

Properties:

TXPxxx4 read (was read by): TX1 Written Text

TX5 Text Recognition

See attached document

TX6 Translitteration

See attached document

TX7 Written Text Segment

Subclass of: TX1 Written Text
Superclass of:
Scope Note: This class describes portions of text considered to be of particular significance by scholars, as witnesses of a certain meaning or bearers of a particular phenomenon relevant to the investigation, study and understanding of a text. Examples of such text portions are columns, fragments, sections, paragraphs, as well as single words or signs, or other components of a written text. To each of these entities can be associated a single production event (TX2) or destruction event (E6), as in the case of letters or words damaged or worn by atmospheric agents or human interventions, as well as specific conditions (E3) for documenting its status during the text recognition process (TX5). The relationship between a written text (TX1) and its components is documented through the TXP4 *has segment* property.

Examples:

- The "INSTINCTV DIVINITATIS" text portion of the inscription (TX1) on the Arch of Constantine (E22), commented by Rodolfo Lanciani in 1892, in his book *Pagan and Christian Rome* (see section 1.3.1).
- The first paragraph of the Darius I's inscription (TX1) in Bagistan.

In First Order Logic:

TX7(x) ⊃ TX1(x)

TX8 Grapheme

See attached document

TX9 Glyph

Subclass of: E25 Man-Made Feature
Superclass of:
Scope Note: This class describes the physical, concrete features traced by a writer, representing the material manifestations of the graphemes needed to codify a linguistic expression. Glyphs are typically observed by the scholars during a text recognition activity (TX5) carried out to decode and recognise the graphemes (TX8) they represent.

Examples:

- The S-shaped feature engraved on the second line of the South inscription on the Arch of Constantine, representing the letter (grapheme) “S” of the Latin writing system used to render the sound of the /s/ phoneme (see section 1.3.1).
- The first feature engraved on the first line of Darius I’s inscription (TX1) in Bagistan, representing the ideal syllabogram of the ancient Persian syllabary, used to render the /da/ syllable.

In First Order Logic:

TX9(x) ⊃ E25(x)

TXxxx New classes

See attached document

TX10 Style

Subclass of: E29 Design or Procedure
Superclass of:
Scope Note: This class describes stylistic variations of a texts, including local script styles (as the Carolingian minuscule for the Latin script) and individual scribal hands. It includes: the general appearance of the script, in terms of general design, aspects related to a bilinear system (i.e,. upper- and lowercases), measures (i.e., large, medium or small), shape and number of strokes forming a character, its order and direction. A style includes also information about ductus (the direction the text), ligatures and *nexi* (i.e., the connection between characters obtained tracing them without detaching the writing instrument from the support and using one or more strokes in common), and the writing angle (i.e., the position the writing instrument is located with respect to the support). The style corresponds to fonts and their variations in modern printing process.

Examples:

- The *Roman square capitals* style, also called *capitalis monumentalis*, or *capitalis quadrata* used to write the inscription on the Arch of Constantine.
- The “Carolingian minuscule” style used in the Carolingian Gospel Book identified as “British Library, Add MS 11848”.

In First Order Logic:

TX10(x) ⊃ E29(x)

5.2 CRMtex - Properties declarations

The properties are comprehensively declared in this section using the following format:

- Property names are presented as headings in bold face, preceded by unique property identifiers;
- The line “Domain:” declares the class for which the property is defined;
- The line “Range:” declares the class to which the property points, or that provides the values for the property;
- The line “Superproperty of:” is a cross-reference to any subproperties the property may have;
- The line “Scope note:” contains the textual definition of the concept the property represents;
- The line “Examples:” contains a bulleted list of examples of instances of this property.

TXP1 used writing system (writing system used for)

Domain: TX2Writing
Range: TX3 Writing System
Subproperty of: P33 used specific technique (was used by)

Quantification: many to many (0,n:0,n)

Scope note: This property identifies the specific instance of TX3 Writing System employed during the writing event (TX2) that led to the creation of a TX1 Written Text.

Examples:

- The Roman stonecutter *used* the Latin *writing system* (TX3) for the engraving (TX2) of the inscription on the Arch of Constantine (TX1) (see section 1.3.1)

- The Greek scribe *used* the Greek *writing system* (TX3) to trace (TX2) in ink the letters that compose the text of the Papyrus of Derveni (TX1).

In First Order Logic:

TXP1(x,y) ⊃ TX2(x)
TXP1(x,y) ⊃ TX3(y)
TXP1(x,y) ⊃ P33(x,y)

TXP2 includes (is included within)

Domain: TX4 Writing Field
Range: TX1 Written Text
Subproperty of: P56 bears feature

Quantification: one to many (0,n:0,1)

Scope note: This property describes the relation existing between a TX1 Written Text and the TX4 Writing Field, specifically created to accommodate the text, within which it is inscribed. This relation becomes quite relevant in the very frequent case where more than a single text is found on different areas of a specific support.

Examples:

- The South framework (TX4) carved by the Roman stonecutter on top of the Arch *includes* the inscription on the South face of the Arch of Constantine (TX1).

In First Order Logic:

TXP2(x,y) ⊃ TX1(x)
TXP2(x,y) ⊃ TX4(y)
TXP2(x,y) ⊃ P56(x,y)

TXP3 rendered (is rendered by)

Domain: TX6 Transliteration
Range: TX5 Text Recognition
Subproperty of P20 had specific purpose (was purpose of)

Quantification: one to one (0,1:1,1)

Scope note: This property describes the close connection for scientific purposes between the TX5 Text Recognition and TX6 Transliteration activities, by outlining, in particular, the accurate observation required by scholars in order to perform a valid transcription or transliteration of a given text.

Examples:

- The transcription (TX6) of the inscription on South Attic of the Arch of Constantine, carried out by Rodolfo Lanciani, *rendered* its reading (TX5) of the same inscription done by him in 1892 (see section 1.3.1).
- The transliteration (TX6) in Latin script of the Mycaenean inscription PY TA 641 by Micheal Ventris.

In First Order Logic:
TXP3(x,y) ⊃ TX5(x)
TXP3(x,y) ⊃ TX6(y)
TXP3(x,y) ⊃ P20(x,y)

TXP4 has segment (is segment of)

Domain: TX1 Written Text
Range: TX7 Written Text Segment
Subproperty of: P46 is composed of (forms part of)

Quantification: one to many (0,n:0,1)

Scope note: This property is intended to correlate a text and the different parts of which a scholar can identify, such as: letters, words, lines, columns, pages, or any other scan that can be made by scholars because considered to have a particular relevance for the investigation of the text itself.

Examples:

- The "INSTINCTV DIVINITATIS" text portion *is segment of* the inscription (TX1) on the Arch of Constantine reported and commented

by Rodolfo Lanciani in 1892 in his book *Pagan and Christian Rome* (see section 1.3.1).

- The phrase "xšâyathiya xšâyâthiânâm" ("King of Kings") recognised by Grotefend in1802 on the Achaemenid inscription from Persia.

In First Order Logic:

TXP4(x,y) ⊃ TX1(x)
TXP4(x,y) ⊃ TX7(y)
TXP4(x,y) ⊃ P46(x,y)

TXP5 wrote (was written by)

Domain: TX2 Writing
Range: TX1 Written Text
Subproperty of P108 has produced (was produced by)

Quantification: one to one (0,1:1,1)

Scope note: This property is used to describe in detail the close relationship between a text and the writing event that led to its production.

Examples:

- The activity (TX2) carried out by the Greek stonecutters *wrote* the Gortyn Law inscription (TX1) on the wall of the Amphitheatre of Gortyn, Crete.

In First Order Logic:

TXP5(x,y) ⊃ TX2(x)
TXP5(x,y) ⊃ TX1(y)
TXP5(x,y) ⊃ P108(x,y)

TXP6 encodes (is encoding of)

Domain: TX3 Writing System
Range: E55 Type (Language)
Subproperty of P2 has type

Quantification: many to many (0,n:0,n)

Scope note This property is used to indicate the language encoded by the TX3 Writing System and used for writing, reading or rendering (i.e. transcribing) a TX1 Written Text.

Examples:

- The Latin alphabet (TX3), used to encode the identical inscriptions (TX1) on the Arch of Constantine, *encodes* the Latin language (E55) used to convey the message of the inscriptions.

In First Order Logic:

TXP6(x,y) ⊃ TX3(x)
TXP6(x,y) ⊃ E33(y)
TXP6(x,y) ⊃ P2(x,y)

TXP7 has item (is item of)

Domain: TX3 Writing System
Range: TX8 Grapheme
Subproperty of P106 is composed of (forms part of)

Quantification: one to one (0,1:1,1)

Scope note: This property is used to state the (conceptual) belonging of a TX8 Grapheme to a given TX3 Writing System.

Examples:

- The Latin alphabet (TX3), used to encode the inscription (TX1) on South face of the Arch of Constantine, *has item* the grapheme <S> (TX8) used in this writing system to represent the /s/ sound.

In First Order Logic:

TXP7(x,y) ⊃ TX3(x)
TXP7(x,y) ⊃ TX8(y)
TXP7(x,y) ⊃ P106(x,y)

TXP8 has component (is component of)

Domain: TX1 Written Text
Range: TX9 Glyph
Subproperty of P46 is composed of (forms part of)

Quantification: one to many (0,n:0,1)

Scope note: This property is used to state the (physical) belonging of a glyph to a given TX1 Written Text.

Examples:

- The inscription (TX1) on South face of the Arch of Constantine, *contains* the S-shaped glyph (TX9) engraved on the second line, representing the letter (TX8) “S” of the Latin writing system (TX3).

In First Order Logic:

TXP8(x,y) ⊃ TX1(x)

TXP8(x,y) ⊃ TX9(y)
TXP8(x,y) ⊃ P46(x,y)

TXP9 is encoded using (was used to encode)

Domain: TX1 Written Text
Range: TX3 Writing System

Quantification: many to many (0,n:0,n)

Scope note: This property has the purpose of directly associating a TX1 Written Text with the TX3 Writing System used for encoding it. It is a shortcut of the more fully articulated path from TX1 Written Text, through TXP5 wrote (was written by), TX2 Writing, TXP1 used writing system (writing system used for) to TX3 Writing System.

Examples:

- The Gortyn Law inscriptions (TX1), engraved on the wall of the Amphitheatre of Gortyn (Crete), *is encoded using* the Greek alphabet (TX3).

In First Order Logic:
TXP9(x,y) ⊃ TX1(x)
TXP9(x,y) ⊃ TX3(y)

TXP10 deciphered text (was deciphered by)

see attached document

NEW PROPERTIES
see attached document

TXP11 translitered (was translitered by)

Domain: TX6 Transliteration
Range: TXxx2 Grapheme sequence
Subproperty of: P16 used specific object (was used for)

Quantification: many to many (0,n:0,n)

Scope note: This property describes the relation between an activity of TX6 Transliteration and the identified sequence of graphemes (TXxx2) represented in an instance of TX1 Written Text.

Examples:

- The transcription (TX6) of the S-shaped feature engraved on the second line of the South inscription of the Arch of Constantine *transcribed* the prototypical letter "S" (TX8) of the Latin writing system (TX3).

In First Order Logic:

TXP11(x,y) ⊃ TX6(x)
TXP11(x,y) ⊃ TX8(y)
TXP11(x,y) ⊃ P16(x,y)

TXP12 has style (is style of)

Domain: TX1 Written Text
Range: TX10 Style
Subproperty of: P33 used specific technique (was used by)

Quantification: many to many (0,n:0,n)

Scope note: This property describes information about the style used for the realization of the written text (TX1). The *TPX12.1 has type* property of *TXP12 has style* allows the nature of the style to be specified reading from domain to range, for example to record the direction, orientation or the linear system of the text.

Examples:

- The Latin text in the Carolingian Gospel Book identified as "British Library, Add MS 11848", has style "Carolingian minuscule"

- The inscription on the Arch of Constantine has ductus (TPX12 has style + TPX12.1 has type "ductus") *dextroverse*

In First Order Logic:

TXP12(x,y) ⊃ TX1(x)
TXP12(x,y) ⊃ TX10(y)
TXP12(x,y) ⊃ P33(x,y)

Properties: TXP12.1 has type: E55 Type

References

Borgwaldt S. R., and T. Joyce (eds), 2013. *Typology of writing systems*, Amsterdam – Philadelphia PA, John Benjamins.

Coulmas F., 1999. *The Blackwell Encyclopedia of Writing Systems*, Oxford – Cambridge MA, Blackwell Publishing.

Daniels, P. T. and W. Bright (eds), 1996. *The World's Writing Systems*, New York – Oxford, Oxford University Press. 1996.

Feldherr, A., and G. Hardy (eds), 2006. *The Oxford History of Historical Writing*. *Volume 1: Beginnings to AD 600*, Oxford, Oxford University Press.

Harris, R., 2005. *Rethinking Writing*, London – New York, Continuum.

Jakobson, R., 1960. "Linguistics and Poetics", in: *Style in Language*, T. Sebeok, ed., Cambridge MA, M.I.T. Press, 350–377.

Lanciani, R., 1892. *The significance of the inscription on the Arch of Constantine,* in *Pagan and Christian Rome*, Houghton, Mifflin and Company, Boston and New York.

Lyons, J., 1972. *Human Language*, in: *Non-Verbal Communication*, R. A. Hinde, ed., Cambridge – New York, Cambridge University Press, 49–85.

Lyons, J., 1977. *Semantics. Volume I*, Cambridge, Cambridge University Press.

Pulgram, E., 1976. *The Typologies of Writing-Systems*, in: *Writing Without Letters*, W. Haas, ed., Manchester, Manchester University Press, 1–28.

Rogers, H., 2005. *Writing Systems: A Linguistic Approach*, Oxford, Blackwell.

Sampson, G., 1985. *Writing Systems: A Linguistic Introduction*, London, Hutchinson (2nd ed. Sheffield, Equinox, 2015).

Sampson, G., 2016. "Typology and the study of writing systems", *Linguistic Typology* 20(3), 561–7.

Saussure, F. de, 1983. *Course in General Linguistics*, Lausanne – Paris, Payot, 1916, translated and annotated by Roy Harris, London, Duckworth, 1983.