Ce qui est écrit et ce qui est parlé. CRMtex for modelling textual entities on the Semantic Web

Achille Felicetti[a] and Francesca Murano[b]
[a]*PIN, VAST-LAB, Prato, Italy*
[b]*Università degli Studi di Firenze, Italy*

Editor(s): Antonis Bikakis, University College London, UK; Beatrice Markhoff, University of Tours, France; Alessandro Mosca, Faculty of Computer Science, Free University of Bozen-Bolzano, IT; Stephane Jean, University of Poitiers - ENSMA, France; Eero Hyvönen, University of Helsinki (HELDIG) and Aalto University, Finland
Solicited review(s): Joini Tuominen, University of Helsinki, Finland; Antonis Bikakis, University College London, UK; Holly Wright, University of York, Great Britian; Oyvind Eide, University of Passau, Germany; One Anonymous reviewer

Abstract. This paper presents the new developments of CRMtex, an ontological model based on CIDOC CRM, created to describe ancient texts and other semiotic features appearing on inscriptions, papyri, manuscripts and other similar supports. The model is also designed to describe in a formal way the phenomena related to the production, use, conservation, study and interpretation of textual entities. CRMtex was originally intended to detect the close relationship linking ancient texts with the physical objects on which they are supported, the tools and writing systems used for their production, and the various scientific investigations and readings carried out on the text by modern scholars. It eventually evolved to provide researchers with the fundamental concepts for the correct and complete rendering of textual objects, the events representing their history and the cultural and social environments in and for which they were created. The full compatibility of CRMtex with the CIDOC CRM ontology and its extensions ensures persistent interoperability of data encoded by means of its entities with other semantic information produced in cultural heritage and digital humanities. The new entities presented in this paper deal more closely with textual and intertextual structures and try to deepen the close relationships existing between fragments of text or sequences of signs and the underlying meaning they were originally intended to convey.

Keywords: Ancient texts, Ontologies, Linguistics, CIDOC CRM, CRMtex

1. Introduction

We have been witnessing of late an intense debate that is animating the world of epigraphists and papyrologists about the need to find or possibly develop conceptual models able to express the complex entities of their domains in a semantically rich encoding, and to establish interoperability of their data with those generated, for example, in the areas of archaeological, historical and linguistic studies. The gigantic integration effort established by Papyri.info [1] and Trismegistos [2] and the various attempts made by projects such as EAGLE [3] to develop a semantic model in the field of epigraphy, testify to a constant and growing interest in the use of advanced and efficient conceptual tools for the generation of standardised, integrated and interoperable information in these disciplines. In the epigraphic world, another important initiative, Epigraphy.info [4], aimed at establishing a collaborative environment for digital epigraphy, is trying to raise awareness in the community of epigraphists about the importance of publishing information in a uniform format and, possibly, in a Semantic Web fashion. This initiative has the merit of having brought all the major players in the epigraphic world around a table and having directed their efforts towards the development of ecosystems in which epigraphic data coming from different sources can be easily retrieved and analysed.

In the same perspective, many European and international initiatives are also focussing their attention on information concerning ancient texts and on the interoperability challenges in which they are involved. The recently completed PARTHENOS project [5] has placed interdisciplinarity at the centre of its activities by designing a system in which historical, archaeological and linguistic data coexist in a single digital environment. ARIADNEplus [6][7], an initiative recently started as a continuation of the first, successful, ARIADNE project, is also attempting to integrate archaeological information and data from other disciplines, with particular regard to the study of archaeological artefacts bearing inscriptions, including amphorae, coins and other similar objects, with the clear intent of creating an interoperable archive based on FAIR principles [8] and international standards. ARIADNEplus is also looking for an ontology or application profile capable of relating textual and archaeological data in a consistent manner. This is one of the gaps that our work aims to fill.

In two of our previous publications [9][10] we tried to give an account of what had been done in the field of epigraphy and what tools had been used to describe, in semantic format, textual entities created in antiquity. In these publications, we also laid the foundations for the definition of a semantic model (CRMepi, later expanded to become CRMtex) centred on the semantic definition of the ancient text and the description of its multifaceted aspects. In the present paper, after a quick review of similar recent initiatives, we present the latest developments of the CRMtex model and the conceptual considerations that underlie its evolution.

2. Looking for new semantic tools

2.1. Ontologies and application profiles: work in progress

Despite the great interest of many scientific communities for the tools proposed by the world of the Semantic Web, it is interesting to note that some wide-ranging initiatives such as EPIDAT [11], the purpose of which is to publish epigraphic data in LOD format, complain of the absence of an ontology able to confer semantic value to epigraphic information. However, an increasing number of activities conducted by groups interested in the subject of ontologies for ancient texts has flourished in recent years.

The ontological approach is also pursued by some major players in the field of epigraphy with various degrees of success: the Epigraphic Database Heidelberg [12], for instance, has released a very basic ontology for the encoding of its vast digital repertoire in Linked Open Data format; however, its model still seems less suitable as a tool for a deep integration.

The Economics and Political Network project (EPNet) [13] is an interesting initiative building an ontological model based on CIDOC CRM to deal with the events and objects connected with the distribution of food in the Roman world. The EPNet ontology looks promising and has already been investigated by the ARIADNEplus project as a prospective part of the application profile for epigraphic data. With the same intent, the Epigraphic Ontology Working Group (EpOnt) [14] is trying to establish an application profile based on concordance of ontologies for recording epigraphic editions. The initiative is interesting and we believe it will produce results very soon.

It should be noted that all these initiatives aim at developing very specific tools for solving the problems concerning the disciplines for which they are conceived. None of them aims to give a common conceptual basis or to look for points of contact, which also exist between these various disciplines, keeping in mind the objective of interoperability.

CRMtex was designed as an ontological model since no existing model is able to investigate thoroughly the textual entities from antiquity, their intrinsic nature as primary sources of knowledge and their link with the archaeological, artistic and historical spheres that make them so precious for the comprehension of the ancient world. In fact, the existing models have been unable to describe the various nuances of a text, from the physical aspect, a set of features created with particular techniques, materials and tools, to the semantic and conceptual aspects, whereby it bears a message that, by means of these same features, is transmitted and disseminated through time and space.

CRMtex was created precisely in order to respond to such needs and to provide tools for modelling textual entities appearing in different contexts by means of standard tools. These justify the use of a tool that certainly requires a considerable investment in expertise, as well as the whole CIDOC CRM ecosystem [15], but which gives its data a richness and a level of interoperability that is difficult to achieve using other systems. The solid foundations of the CIDOC CRM, on which it is built, already provide the necessary top-level classes and properties to model objects, events, actors, spatial and temporal entities in a standard way, leaving full freedom to its classes and properties to focus on the issues concerning text and its material production in antiquity [16]. This same compatibility is what allows the model to define textual entities as *tesserae* of those great knowledge puzzles that are semantic graphs, providing papyro logists, epigraphists and those who deal with graffiti or ancient and modern manuscripts [17] with a tool to encode knowledge natively in interoperable and reusable formats.

2.2. EpiDoc: a de facto standard for ancient texts

It should be emphasised that epigraphists and papyrologists have long since chosen TEI EpiDoc [18] as their own metadata standard, as this tool is extremely versatile for representing texts and the phenomena that typically characterise them, with particular attention to the needs of a rich and well-rendered visualisation.

EpiDoc, which for the treatment of the text is based on the Leiden conventions [19], provides a series of tags for detecting specific elements, since the text itself may contain semantically relevant information that needs to be captured in some way [20]. Interesting examples in this sense are the tags that identify temporal entities, actors and place names, which give EpiDoc the ability to bind external semantic elements starting from identifiable textual fragments.

Nevertheless, it should also be noted that EpiDoc does not offer the typical descriptive tools used by ontologies to capture the conceptual *nuances* of the text as a material phenomenon framed in time and space, and to define metadata that can describe its structure, history and the events and people who determined its existence and life. In this context, it becomes essential to use models that can put the TEI ecosystem in touch with the universe of the ontologies and to act as a link between these different worlds. In the field of numismatics, for instance, the Nomisma.org project [21] has successfully attempted to act as a link between different numismatic resources by integrating specific vocabularies, models and ontologies. CRMtex tries to propose a similar solution by establishing a solid conceptual basis for bridging knowledge of different types and implementing interoperability for textual data in an effective way.

3. CRMtex: an ontology for ancient texts

The need to create a new ontology for ancient texts started from the assumption that, unlike printed texts, non-mechanised written texts (including inscriptions, papyri and manuscripts) have specific features that must be taken into account for their study.

We have based our model on the solid foundations of CIDOC CRM because it constitutes one of the most widely used ontologies in the field of Cultural Heritage. In its core version, CIDOC CRM already provides most of the entities necessary to model common elements such as actors, objects, places, events and their mutual interrelations on a chronological basis.

CRMtex has its foundation in the semiotic aspects of language and text [22]; the core concept of our model is therefore the notion of “text” as the product of a semiotic process involving an encoding (“writing”) and a decoding (“reading”) process. Writing is in turn a particularly sophisticated human technology allowing the encoding of a linguistic message through a series of signs specifically selected for this purpose [23][24].

Investigating in detail the close relationship that links the text with the writing event, some considerations to clarify its nature need to be set forth.

Although every speech can be transposed into an equivalent written message, and *vice versa*, speech has a priority over writing in at least four respects: phylogenetically, ontogenetically, functionally and structurally [25]. In fact, languages are all spoken but not necessarily written; every human being spontaneously learns to speak naturally, the ability to write coming only later and through specific training; the spoken language is employed in a wider and differentiated range of uses and functions; writing originated as a representation of speech. Indeed, according to Ferdinand de Saussure [26], “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”.

In this semiotic perspective, it is worth considering that even in writing, as in analysis of the linguistic system, it is necessary to distinguish the concrete level of the personal execution (i.e., the real act of tracing signs on a surface) from the abstract level, to which all the single occurrences must be taken back, on the basis of a principle of identity or sameness (e.g., identification of an “A”, independently from the peculiar shape someone may give to it).

Thus, a "text" is constituted by a number of signs physically traced (i.e., *written*) on a specific support and intended to encode a linguistic expression.

Because of their non-mechanised origin, ancient texts are unique and unrepeatable entities; in addition, along with their support, they form an inextricably linked, unique object of study. Thus from a conceptual point of view, whether it is painted, written in ink or engraved, a text preserves its physical nature, which is a feature deriving its existence from its strict dependence on the support on which it is located.

CRMtex provides specific entities to describe all these phenomena and, being an extension, takes advantage of the power of CIDOC CRM and its other extensions (e.g., CRMsci, the scientific observation model [27][28], CRMarchaeo, a model for archaeological excavation documentation [29][30]) to describe general, non-textual information (i.e., actors, places, objects, temporal entities, observations, archaeological contexts and so forth). In its current version (v1.0), CRMtex is composed of 9 classes and 11 properties; all of them are defined as subclasses and sub-properties of CIDOC CRM classes and properties. A brief description of each of them is provided below. The full documentation, with all the scope notes and examples, can be found on the CIDOC CRM Extensions web pages [31]. An RDFS version of the model is also provided [32].

3.1. CRMtex classes

CRMtex in its previous version provided a set of 6 classes, which we have covered extensively in [9] and [10]. In chapter 4 of this paper we present the new classes (*TX7*, *TX8*, *TX9*) that have been incorporated in the new 1.0 version, currently under evaluation and approval by the CIDOC CRM Special Interest Group. CRMtex classes include:

TX1 Written Text, a subclass of CIDOC CRM *E18 Physical Feature*, is intended to describe the physical signs composing a text that is engraved or incorporated on or into some kind of physical support, having semiotic significance and the intentional purpose of conveying a linguistic message.

TX2 Writing, a subclass of *E12 Production*, indicates the activity of creating permanent marks on a physical support using various techniques (painting, sculpture, etc.), by means of specific tools. The *TXP5 was written by* property (a subproperty of *P108 was produced by*) is used to render in a clear way the link between the text (*TX1*) and its production (*TX2*).

TX3 Writing System, a subclass of *E29 Design or Procedure*, represents the conventional set of signs and the related rules used to codify and represent (i.e., to *write*) utterances meant to be recovered at a distance of time and/or space by those who have knowledge of the same code (i.e., the same linguistic system). The *TXP9 is encoded by* property provides a direct link between the text (*TX1*) and the writing system (*TX3),* thus offering the possibility to describe this relation in more generic terms.

TX4 Writing Field, a subclass of *E25 Man-Made Feature*, represents the portion of the physical support 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, especially in epigraphy, in which a specific element called the "epigraphic field" has been defined by the discipline itself. Its importance is also evident in papyrology and codicology, where a clear distinction between the area(s) containing the written text and empty parts of the support (margins, *intercolumnia*, etc.) is significant for the definition of the styles and periods of a document.

In addition to dealing with a text as an object, our model also focusses on the research procedures, and provides classes and relationships to describe the typical operations that scholars from different disciplines perform in order to gain knowledge about textual entities. It is evident, in this perspective, that the study of ancient texts typically starts from the analysis of the physical characteristics of the text itself before moving to the investigation of their archaeological, palaeographic, linguistic and historical features. In this regard, we have defined the following classes:

TX5 Reading, a subclass of the CRMsci *S4 Observation* class, refers to the semiotic procedure of decoding (and therefore understanding) a written text. This procedure can be carried out for scientific purposes, in order to analyse and study the text according to different disciplinary perspectives. The reading activity is thus intended as a specific observation (*S4*) in which the decoding of the signs is performed, i.e., the linguistic value is recognised and the message is understood.

TX6 Transcription, a subclass of *E7 Activity*, refers to the activity of re-writing the text by an editor. This operation could involve a writing system (*TX3*) different from that of the original text, implying a transposition of the sounds of a language from one writing system to another (e.g., Latin letters to render a Mycenaean text).

TX7 Written Text Segment, a subclass of *TX1 Written Text,* is intended to identify 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 the text.

TX8 Grapheme is a subclass of *E90 Symbolic Object*, used to represent the abstract units with distinctive value in a given writing system. A grapheme is a character or sequence of characters that functions as a distinct unit within an orthography. It may be a single character, a multigraph or a diacritic but, in all cases, graphemes are defined in relation to the particular orthography.

TX9 Glyph, a subclass of *E25 Man-Made Feature,* represents the concrete manifestation of single signs traced by the writer while codifying a linguistic expression. Glyphs are typically observed by the scholar during a reading activity (*TX5*) carried out to decode and recognise the graphemes (*TX8*) they represent.

3.2. CRMtex properties

CRMtex also provides adequate properties to link instances of its classes. A list of the properties is provided below. The use of the new properties in version 1.0 (*TXP4, TXP7, TXP8* and *TXP11*) is investigated in Chapter 4 together with the description of the new classes.

TXP1 used writing system (writing system used for), a subproperty of *P33 used specific technique (was used by)*, is intended to identify the specific instance of *TX3 Writing System* employed during the writing event that led to the creation of a *TX1 Written Text*.

TXP2 includes (is included within) is a subproperty of *P56 bears feature* intended to describe the relation existing between instances of *TX1 Written Text* and of the *TX4 Writing Field* specifically created to accommodate them.

TXP3 rendered (is rendered by), is a subproperty of *P20 had specific purpose (was purpose of)*, which links an instance of *TX6 Transcription* with the *TX5 Reading* activity carried out by scholars to establish an accurate rendering of the text being investigated.

TXP4 has segment (is segment of) is a subproperty of *P46 is composed of (forms part of)*, intended to correlate a text (*TX1*) and the different parts (*TX7*) of it that a scholar can identify, such as letters, words, lines, columns, pages, or any other scholarly relevant segment.

TXP5 wrote (was written by) is a subproperty of *P108 has produced (was produced by)*, used to describe in detail the close relationship between a text (*TX1*) and the writing event (*TX2*) that led to its production.

TXP6 encodes (is encoding of) is a subproperty of *P2 has type*, used to indicate the language (*E33*) encoded by the *TX3 Writing System* and used for writing, reading or rendering (i.e., transcribing) a *TX1 Written Text*.

TXP7 has item (is item of) is a subproperty of *P106 is composed of (forms part of)*, used to state the (conceptual) belonging of a *TX8 Grapheme* to a given *TX3 Writing System*.

TXP8 has component (is component of) is a subproperty of *P46 is composed of (forms part of),* used to state the (physical) belonging of a *TX9 Glyph* to a given *TX1 Written Text*.

TXP9 is encoded using (was used to encode) has the purpose of directly associating a *TX1 Written Text* with the *TX3 Writing System* from which have been taken the signs that have been used for its writing and have been incorporated in the text.

TXP10 read (was read by), a subproperty of the CRMsci *O8 observed (was observed by)* property, links an instance of *TX1 Written Text* with a *TX5 Reading* event carried out to investigate its intrinsic characteristics and to perform its decoding.

TXP11 transcribed (was transcribed by), a subproperty of *P16 used specific object (was used for)*, highlights the specific way in which an activity of *TX6 Transcription* results in the rendering of the specific *TX8 Grapheme(s)* of which an instance of *TX1 Written Text* is composed.

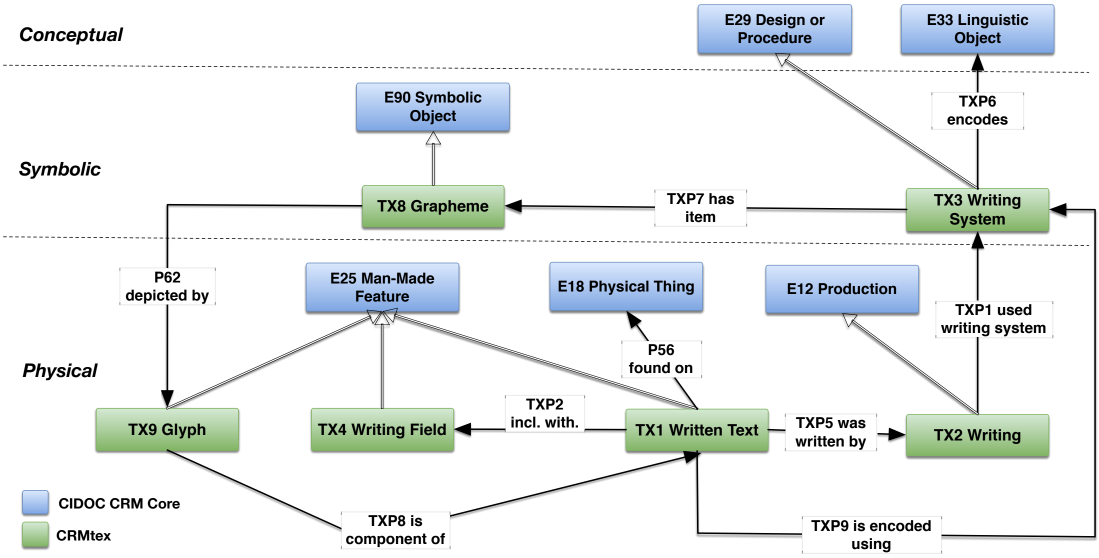

Figure 1. General overview of the new CRMtex model.

The overall scheme of CRMtex classes and properties is presented in Figures 1 and 2. In particular, Figure 1 presents the modelling of the text and its production, considering the three different encoding levels: the level concerning the written text, i.e., the glyphs (physical level); the level concerning the graphemes and the other symbolic entities encoded at the time of writing (symbolic level); the level concerning the ideas and concepts intended to be expressed and communicated by the text (conceptual level). Figure 2 offers the point of view of the investigation of the text with the entities related to its reading (i.e., the accurate observation of its physical features) and its transcription.

4. Between Semiotics and Linguistics: new entities in CRMtex

4.1. Written Text Segments

In designing the new entities of our model, we began by thoroughly investigating the interconnections existing between the text and its various components. We have also tried to establish a complete chain of connections to link these components and the whole text with the linguistic level they encode. Some elements have proved to be absolutely essential for this purpose. Concerning the reading process (i.e., the decoding of the text), and therefore the investigation of the text by scholars, one has shown particular importance, namely, the text segment element.

We therefore introduced the new *TX7 Written Text Segment* class, a subclass of *TX1 Written Text*, intended to identify portions of text considered to be of particular significance as witnesses of a certain meaning, or bearers of special phenomena relevant to the investigation, study and understanding of the text (see Figure 2). Examples of text portions are text columns, text fragments, sections, paragraphs, single words or letters etc.

Scholars of different disciplines need to identify such segments, based on the requirements of their study, and to focus their attention on them in order to describe their physical properties (form, layout etc.), to verify their legibility or to identify particular phenomena (e.g., linguistic or palaeographic aspects) that are connected to them. When modelling, it is important to define unambiguously such segments and their relationship with the text in its entirety, so as to be able to assign specific properties to the individual segments, independently of the text as a whole. Particular production (*TX2*) or destruction (*E6*) events can be associated to each fragment as in the case of letters or words damaged or worn by atmospheric agents or human interventions. Specifications about conditions (*E3*) for documenting the status of the text during the observation process (*S4*) can be easily stated as well. This allows scholars to document different events for the investigated segments in a more precise and punctual way and to assign observations and interpretations to them.

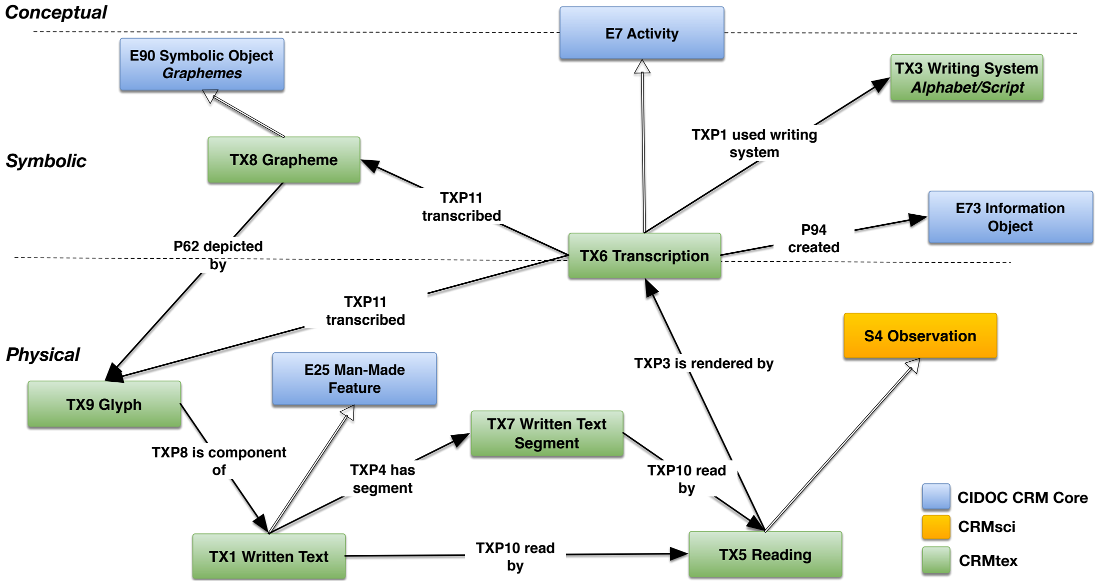

Figure 2. Reading and Transcription activities in CRMtex.

The relationship between a written text (*TX1*) and its components is documented through the *TXP4 has segment* property (see Figure 3).

4.2. Glyphs and Graphemes

The physical signs composing a *TX1 Written Text* constitute the material manifestations (glyphs) of writing system units, i.e., the graphemes, which are the minimal functional distinctive units of writing. Ernst Pulgram has stated that "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 choose as his basis" [33]. For instance, in a Latin inscription, single letters of the alphabet (glyphs) represent graphemes, a grapheme corresponding to a letter only in alphabetic system of writing. In Mycenaean Linear B inscriptions and in Old Persian cuneiform inscriptions, glyphs represent syllabograms (graphemes representing a syllable, not a single sound); in an Egyptian hieroglyphic text, glyphs represent syllabic, alphabetic and also ideographic elements, i.e., elements standing for lexical/semantic units.

Phonographic writing systems [34][35] represent phonological units of one size or another, but the 1:1 correspondence between sound (phoneme, syllables *etc.*) and sign (grapheme) is lost in diachrony, obscured by spelling conventions and phonetic changes to which linguistic systems are subjected in the course of history. Think of the spelling discrepancies in English between writing and reading: for example, the <i> grapheme stands for various phonemes: /ɪ/ (as in *him*), /ʌɪ/ (as in *time*), /i/ (as in *police*), /a/ (as in *timbre*); *vice versa*, the /f/ phoneme can be represented by <f> (as in *film*), <ph> (as in *philology*), <gh> (as in *enough*).

Concerning the message retrieval, reading the written message presupposes the ability to read the language of the writer since each grapheme is bound to a given linguistic unit of specific languages.

In this view, the model provides two new classes to represent the units the scholars deal with: *TX9 Glyph*, a subclass of *E25 Man-made Feature*, and *TX8 Grapheme,* a subclass of *E90 Symbolic Object* (see Figure 1). Specific properties are used to settle the strict correspondence between graphemes and glyphs and their typical parthood relationships such as the *TXP7 has part*, used to state the (conceptual) belonging of a grapheme to a given writing system, and the *TXP8 is contained in*, which is used to state the (physical) belonging of a glyph to a given text or segment of text. The *TXP9 is encoded using* is used to state in a more general way that the graphemes used to compose the text (*TX1*) belong to a given writing system (*TX3*).

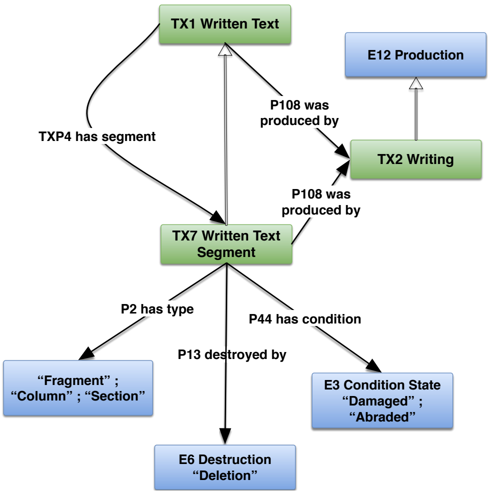

Figure 3. Written Text and Written Text Segment in CRMtex.

Moving to the level of the linguistic sounds, it will be the decoders (readers, including scholars) who, from time to time, on the basis of their knowledge of the linguistic system, will attribute to each sign or group of signs the adequate phonetic value, also doing so on the basis of spelling conventions present in a given graphic system at a given historical moment, since the orthographic rules can change over time, even if less quickly than the linguistic system does. The ontological description of the link between linguistic and graphic units is under preparation by the authors.

5. Application scenarios

5.1. CRMtex and EpiDoc

In designing our model, we have always tried to maintain the compatibility of our entities with those of EpiDoc. A natural compatibility with our model obviously exists for the information present in the `teiHeader` section of EpiDoc, containing the metadata of documents and corpora described in the files. We have also made CRMtex classes and properties suitable to describe, in addition, peculiar phenomena of the text and its conditions for which EpiDoc tags are usually used. For instance, in presence of characters erased in antiquity but still legible in a more or less clear way, EpiDoc employs the following syntax:

```
  <del rend="erasure" unit="character"
quantity="3">abc</del>
```

The same information can be expressed in CRMtex by combining the *P43 has dimension* and the *P44 has condition* properties (plus the “erasure” concept from the Getty AAT Thesaurus [36]) in the following way:

```
<http://crm.tx/text102/fragment5>
   a   crmtex:TX7_Written_Text_Segment ;
   crm:P44_has_condition   aat:300053088 ;
   crm:P43_has_dimension   <frg5_dim> ;

aat:300053088
   a crm:E3_Condition_State , skos:Concept
   rdfs:label  “erasure”@en

<frg5_dim>
   a   crm:E54_Dimension ;
   crm:P90_has_value   “3” ;
   crm:P91_has_unit    “character”
   crm:P3_has_note     “abc”
```

More details can be specified for each character, if necessary, by instantiating a *TX9 Glyph* class in order, for example, to describe the specific circumstances under which they were damaged.

Diversely, an erasure indicating a text that is lost and is thus illegible, encoded in TEI EpiDoc (XML) as:

```
<del rend="erasure">
 <gap reason="lost" quantity="4"
  unit="character"/>
</del>
```

implies, according to CRMtex, the use of an *E6 Destruction* class, indicating an event that took out of existence (and out of the support) the signs of the original text, rendering them unrecoverable. CRMtex describes such a phenomenon in the following way:

```
<http://crm.tx/text102/fragment13>
   a   crmtex:TX7_Written_Text_Segment ;
   crm: P13i_destroyed_by   <frg13_dest> ;
   crm:P44_has_condition   aat:300053088 ;
   crm:P43_has_dimension   <frg13_dim> .

aat:300053088
   a crm:E3_Condition_State , skos:Concept
   rdfs:label  “erasure”@en

<frg13_dest>
```

```
   a   crm:E6_Destruction .

<frg13_dim>
   a   crm:E54_Dimension ;
   crm:P90_has_value   "4" ;
   crm:P91_has_unit   "character" .
```

RDF notation is certainly less concise than that provided by EpiDoc, but it is also more expressive and able to specify the historical or environmental circumstances that determined a particular condition of the text, to provide a deeper level of standardisation in the formalisation of knowledge (for example, through the use of thesauri) and to link external relevant entities for implementing information enrichment in multiple stages, including after the initial encoding. The use of the *E54 Dimension* class, for instance, could offer the opportunity to specify historical information about the events, times and circumstances in which the text was lost, when such data are retrieved in other datasets or come to light during the research work. What is more, the destruction event that produced the erasure in the example above could be identified as an historical event documented by other sources, allowing this fragment of knowledge to become part of a larger knowledge graph.

The complexity of encoding in RDF responds to the need to describe in detail all the events involved in the production of the text to be encoded. It is clear, however, that it is not necessary to choose between EpiDoc and CRMtex since the two tools respond to different research needs. CRMtex is not aimed at the digital edition of the text, for which the EpiDoc XML encoding already works well, but at capturing and describing knowledge related to the text itself in a holistic perspective. However, the two models can be used in synergy to create richer metadata and build more structured and complete information from both a descriptive and semantic point of view, thus fostering interoperability of textual information in the typical integrated scenarios of the Semantic Web.

5.2. The inscription on the Arch of Constantine

To illustrate the features of the new version of the CRMtex, we propose an epigraphic example: the inscriptions on the Arch of Constantine, one of the most famous ancient monuments in Rome.

Other examples of the application of our model are illustrated in our previous publications: in [9] it is applied to the encoding of an inscription in Oscan, a language of fragmentary attestation; in [10] is used in a different field of application, that of papyrology, for the encoding of the Derveni papyrus.

The Arch, 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 era) 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 [37], originally inlaid with gilded bronze letters, explaining the reason for its construction.

The bronze letters are now lost and only the large cuttings remain in the marble, in which the bronze letters were fixed. The text is repeated, identically, on the South and North faces of the arch. A transcription and a translation in English of the same inscription is presented below.

Transcription of the inscription:

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

Translation of the inscription:

"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".

From the CIDOC CRM point of view, the Arch is an archaeological object (i.e., an *E22 Man-made Object*) made of marble, mainly intended to commemorate the emperor. Two distinct writing events (*TX2*) can be assigned to the inscriptions, to describe the different production phases of each of them and to distinguish them from the production of the monument.

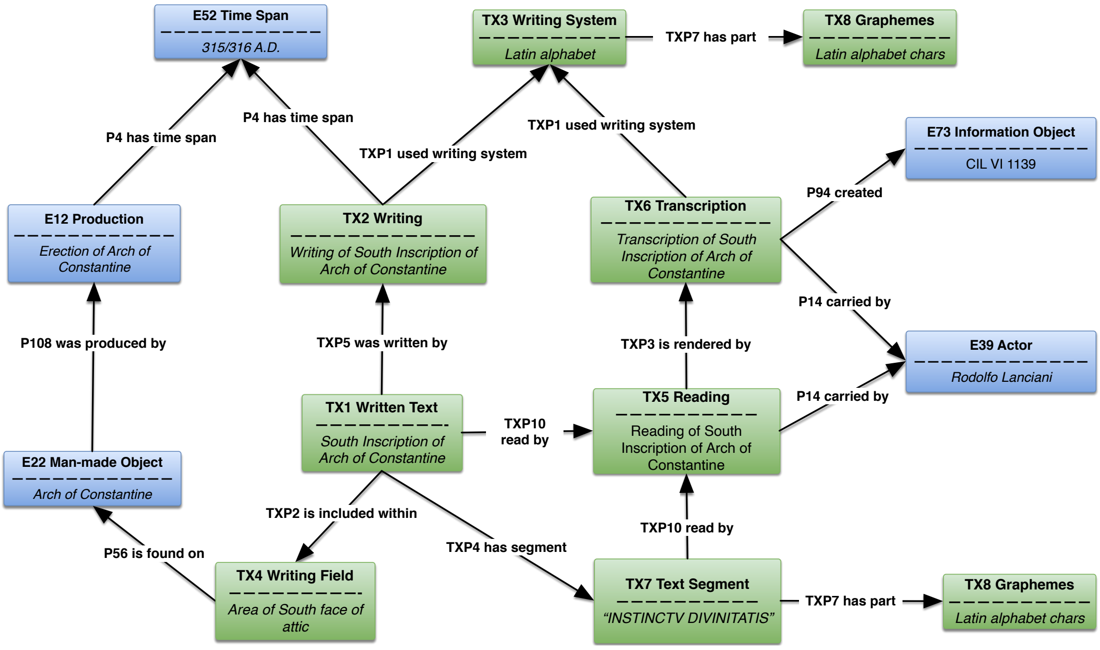

Figure 4. CRMtex modelling of the inscription on the South side of the Arch of Constantine.

CRMtex can be used to describe the 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.

An instance of the *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 *TX4*s instances.

The linguistic message conveyed by the inscriptions (*E90 Symbolic Object*) is encoded by means of the writing system this language uses. It follows that the *TX1 Written Text* class is the concrete graphical manifestation (i.e., the signs – in this case the Latin letters – that 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.

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. In addition, 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, aimed at studying and processing the inscribed text, CRMtex provides specific classes and properties. The transcription of the text(s) present in *Corpus Inscriptionum Latinarum*, 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 [38], can be documented using the *TX5 Reading* class, underlying the scientific nature of the investigation. Reading (*TX5*) and transcription (*TX6*) activities can be related via the *P20 has specific purpose* property, inherited by CIDOC CRM core.

The *TX7 Written Text Segment* class can be used to highlight portions of text on which the study focusses, on which peculiar phenomena appear or from which special meanings are derived. Rodolfo Lanciani, for instance, investigated the "INSTINCTV DIVINITATIS" phrase, advancing an hypothesis on its real meaning in the framework of the message Constantine intended to transmit to the inhabitants of Roman Empire (both Christians and Pagans). Figure 4 shows only a CRMtex general rendering of one of the inscriptions on the Arch of Constantine: more detailed descriptions of the text and the way it was investigated, thus expanding the semantic knowledge graph concerning this monument, can be defined when required.

The *TX9 Glyph* class in combination with the *TXP3 is rendered by* property, for instance, can be used to model one of the typical phenomena of Roman epigraphy, i.e., the use of specific signs as abbreviations, also present in this text (e.g., "S" for "SENATUS", "P" for "POPULUS" etc.). Associating abbreviation expansions to these glyphs would be ideal to document the choices made by scholars for resolving abbreviations during the transcription phases. The considerations that motivated these interpretative choices can be expressed by means of CRMinf [31], the extension of the CIDOC CRM developed to support argumentations and to document inferences and the formulation of hypotheses.

6. Conclusions

CRMtex was developed by adopting the best modelling principles of the ontological world and the fundamental paradigms of linguistic research: this makes it a tool capable of conferring ontological value to textual entities, offering innumerable benefits for research in many humanistic disciplines. The possibility to provide representation of cultural data on the Semantic Web, to publish them in standard formats (such as LOD) and to make them easily available, interoperable and reusable in an infinite number of contexts, certainly represents one of the most relevant features of the model.

The native ability of CRMtex to describe relationships between text and artefacts by efficiently placing the text in the context of the life and history of ancient objects, also makes it ideal for employment in projects like ARIADNEplus or in initiatives like Epigraphy.org. The perfect compatibility with EP-Net, the model used by some ARIADNEplus partners to codify epigraphic information, will foster the possibility for CRMtex to become part of the Application Profile for epigraphy under definition within this project.

Nevertheless, a lot of work still remains to be done for the ontology to reach its maturity.

In 2018 CRMtex was accepted as part of the CIDOC CRM family [39], thus becoming a new tile of the CIDOC CRM mosaic of models. A process of fine tuning to make CRMtex perfectly integrated and consistent with the other extensions of this ecosystem is already under way. In particular, we will need to plan harmonisation with CRMinf [40], the importance of which we have already stressed for the interpretation of the text (see Chapter 5), and with FRBRoo [41], a CIDOC CRM compatible model aimed at representing the semantics of bibliographic information. Many FRBRoo classes (such as the *F2 Expression*, *F12 Nomen* and *F23 Expression Fragment*) actually present interesting points of contact with CRMtex and could form the basis for the creation of a more complex (but more complete) ontological instrument for the effective modelling of (ancient and modern) textual entities.

Despite being a relatively new model that is still under development, CRMtex is already used in many contexts where the definition of textual entities from the ancient world is fundamental, especially in the Cultural Heritage field. Thus, CRMtex is employed by various initiatives of far-reaching national and international scope for this aim. CRMtex has been selected as the ontological model for the project "Aggressive magic in the ancient world: lexicon and formulae of Greek texts" of the University of Florence (Italy) [42], focussing on the study of Greek curse tablets and magical papyri. In the ARIADNEplus framework, it has been chosen as a candidate for the encoding and integration of inscription and graffiti data in the semantic infrastructure that the project is building. CRMtex has also been selected among the basic models for the ontology in the process of definition by the community of epigraphists in the framework of the epigraphy.info initiative for the interoperability of epigraphic data.

The model is constantly expanding and is oriented towards the deepening of linguistic aspects that could enhance its skills and foster its use in other disciplines. Thus, among future activities, we aim to investigate the close correlation of graphemes with the linguistic units (such as phonemes) of which they are conceptual representations and the way in which, through phonemes, the thought of the speaker (and therefore of the writer) materialises in the form of linguistic expressions to become text. We shall then extend CRMtex with new entities that are suitable to describe such complex linguistic phenomena.

References

[1] Papyri.info Initiative, http://papyri.info.

[2] Trismegistos Portal, http://www.trismegistos.org.

[3] EAGLE Project, https://www.eagle-network.eu.

[4] Epigraphy.info Initiative, http://epigraphy.info.

[5] PARTHENOS Project, http://www.parthenos-project.eu.

[6] ARIADNEplus Infrastructure, https://ariadne-infrastructure.eu.

[7] C. Meghini, R. Scopigno, J. Richards, H. Wright, G. Geser, S. Cuy, J. Fihn, B. Fanini, H. Hollander, F. Niccolucci, A. Felicetti, P. Ronzino, F. Nurra, C. Papatheodorou, D. Gavrilis, M. Theodoridou, M. Doerr, D. Tudhope, C. Binding, A. Vlachidis, ARIADNE: A Research Infrastructure for Archaeology, ACM Journal on Computing and Cultural Heritage 10/3 (2017), https://doi.org/10.1145/3064527.

[8] FAIR Guiding Principles for scientific data management and stewardship', https://www.go-fair.org/fair-principles/.

[9] A. Felicetti, F. Murano, P. Ronzino, F. Niccolucci, CIDOC CRM and Epigraphy: A Hermeneutic Challenge, in: Extending, Mapping and Focusing the CIDOC CRM. Proceedings of CRMEX 2015, P. Ronzino, ed., CEUR-WS.org, 2016, 55–68 (http://CEUR-WS.org/Vol-1656).

[10] A. Felicetti, F. Murano, *Scripta Manent*: A CIDOC CRM Semiotic Reading of Ancient Texts, International Journal on Digital Libraries 18/4 (2017), 263–270.

[11] T. Kollatz, EPIDAT - Research Platform for Jewish Epigraphy, in Crossing Experiences in Digital Epigraphy. From Practice to Discipline, A. De Santis, I. Rossi, eds., Warsaw – Berlin, De Gruyter, 2018, 231–239.

[12] Epigraphic Database Heidelberg Ontology, https://edh-www.adw.uni-heidelberg.de/edh/ontology.

[13] D. Calvanese, P. Liuzzo, A. Mosca, J. Remesal, M. Rezk, G. Rull, Ontology-Based Data Integration in EPNet: Production and Distribution of Food During the Roman Empire, Engineering Applications of Artificial Intelligence 51 (2016), 212–229.

[14] EpOnt (Epigraphic Ontology Working Group): https://groups.google.com/forum/#!forum/epont.

[15] CIDOC CRM v.6.2.3, http://www.cidoc-crm.org/Version/version-6.2.3.

[16] G. Bruseker, N. Carboni, A. Guillem, Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM, in: Heritage and Archaeology in the Digital Age. Quantitative Methods in the Humanities and Social Sciences, M. Vincent, V. López-Menchero Bendicho, M. Ioannides, T. Levy, eds., Springer, Cham, 2017, https://doi.org/10.1007/978-3-319-65370-9_6.

[17] A. Felicetti, F. Murano, CRMtex: semantic and semiotic strategies for the encoding of hand-written documents, in: Recueil des résumés, Atelier "Les manuscrits de Saussure, parmi d'autres" - Colloque International "Le Cours de linguistique générale. 1916-2016. L'émergence", Genève 9-14 jan. 2017, 120. https://www.clg2016.org/pdf/clg2016-recueil-des-resumes.pdf.

[18] EpiDoc: Epigraphic Documents in TEI XML, http://sourceforge.net/p/epidoc/wiki/Home/.

[19] H. Krummrey, S. Panciera, Criteri di edizione e segni diacritici, Tituli 2 (1980), 205–215.

[20] T. Elliott et al., All Transcription Guidelines, in: EpiDoc Guidelines, (2019), http://www.stoa.org/epidoc/gl/dev/app-alltrans.html.

[21] Nomisma.org Collaborative Project, http://nomisma.org.

[22] P. Cobley (ed.), The Routledge Companion to Semiotics, London – New York, Routledge, 2001.

[23] R. Harris, La Sémiologie de l'écriture, Paris, CNRS, 1994.

[24] R. Harris, Signs, Language and Communication, London, Routledge, 1996.

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

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

[27] CRMSci: Scientific Observation Model, http://www.cidoc-crm.org/crmsci/.

[28] F. Niccolucci, Documenting archaeological science with CIDOC CRM, International Journal of Digital Libraries 18 (2017), 223–231. https://doi.org/10.1007/s00799-016-0199-x.

[29] CRMarchaeo: Excavation Model, http://www.cidoc-crm.org/crmarchaeo/.

[30] M. Doerr, A. Felicetti, S. Hermon, Definition of the CRMarchaeo, an extension of CIDOC CRM to support the archaeological excavation process, Technical report 1.4.8.

[31] CIDOC CRMtex, Model for the study of ancient texts, http://www.cidoc-crm.org/crmtex/.

[32] CRMtex Model RDF Version, https://github.com/Akillus/CRMtex

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

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

[35] H. Rogers H., Writing Systems: A Linguistic Approach, Oxford, Blackwell, 2005.

[36] Getty Art & Architecture Thesaurus, https://www.getty.edu/research/tools/vocabularies/aat/

[37] CIL VI 1139.

[38] R. Lanciani, Pagan and Christian Rome, Boston – New York, Houghton, Mifflin and Company, 1892.

[39] CIDOC CRM Compatible Models and Collaborations, http://www.cidoc-crm.org/collaborations.

[40] CRMinf: Argumentation Model, http://www.cidoc-crm.org/crminf.

[41] FRBRoo: Functional Requirements for Bibliographic Records, http://www.cidoc-crm.org/frbroo/.

[42] Progetto "La magia aggressiva nel mondo antico: lessico e formulario dei testi in greco", https://www.progettomagia.unifi.it/.