Umanistica Digitale – ISSN: 2532-8816 – n. 11, 2021
A. Felicetti, F. Murano – *La modellazione semantica delle entità testuali*
DOI: http://doi.org/10.6092/issn.2532-8816/13674

La modellazione semantica delle entità testuali

Il modello CRMtex e la descrizione ontologica dei testi antichi

[1]Achille Felicetti, [2]Francesca Murano

Università di Firenze, Firenze, Italia
[1]achille.felicetti@pin.unifi.it
[2]francesca.murano@unifi.it

Abstract

Questo articolo presenta CRMtex, un modello ontologico basato su CIDOC CRM sviluppato dal 2015 per supportare lo studio di documenti antichi. Il modello ha lo scopo di identificare entità testuali rilevanti e modellare il processo scientifico relativo allo studio dei testi antichi e delle loro caratteristiche al fine di favorire l'integrazione con altri campi di ricerca relativi al patrimonio culturale. CRMtex è in grado di identificare e definire in modo chiaro e univoco le principali entità coinvolte nello studio e nell'edizione di testi manoscritti e di descriverle mediante opportuni strumenti ontologici in una prospettiva multidisciplinare. Il modello CRMtex fornisce inoltre strumenti per gestire questo tipo di complessità definendo classi e proprietà per descrivere un testo manoscritto in tutti i suoi aspetti, dalla sua creazione (e/o distruzione) nel passato, fino alla sua attuale conservazione, indagine e studio da parte degli studiosi, compresa la sua trascrizione, traduzione, interpretazione e pubblicazione. La piena compatibilità di CRMtex con l'ontologia CIDOC CRM e le sue estensioni garantisce l'interoperabilità persistente dei dati codificati per mezzo delle sue entità con altre informazioni semantiche prodotte nell'ambito dei Beni Culturali e delle Digital Humanities.

This paper presents CRMtex, an ontological model based on CIDOC CRM developed since 2015 to support the study of ancient documents. The model is intended to identify relevant textual entities and to model the scientific process related to the investigation of ancient texts and their features in order to foster integration with other Cultural Heritage research fields. CRMtex is able to identify and define in a clear and unambiguous way the main entities involved in the study and edition of ancient handwritten texts and to describe them by means of appropriate ontological instruments in a multidisciplinary perspective. The CRMtex model also provides tools for managing this kind of complexity by defining classes and properties for describing a handwritten text in all its aspects, from its creation (and/or destruction) in the past, down to its present conservation, investigation and study by scholars, including its transcription, translation, interpretation and publication. 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.

Introduzione e stato dell'arte

Rispetto ai testi moderni prodotti con tecniche meccanizzate, le testimonianze scritte che sono arrivate fino a noi dall'antichità si caratterizzano per la loro unicità, dal momento che ognuna di esse è frutto di una performance individuale, unica e irripetibile. Come per ogni attività umana, anche la scrittura avviene *hic et nunc*. Infatti, anche nel caso di testi scritti dalla stessa persona su supporti identici e con identica tecnica, i segni tracciati risultano comunque sempre differenti e distinguibili; al contrario, a partire dall'invenzione della stampa industriale, le moderne copie stampate di libri e documenti possono essere considerati nella sostanza indistinguibili da un esemplare all'altro, poiché i caratteri sono incisi da un'identica matrice. Anche i primi testi stampati, creati prima della rivoluzione industriale, sono esemplari unici, poiché sono stati prodotti con meccanismi artigianali e non industriali, quali ad es. l'uso di caratteri tipografici realizzati a mano. Questo aspetto di unicità si ravvisa, per i documenti scritti antichi, anche laddove questi testi siano duplicati (per es. la doppia iscrizione dell'Arco di Costantino) o prodotti in serie attraverso processi meccanizzati (per esempio i bolli sui laterizi o le legende monetali), dal momento che ogni singolo testo possiede una propria storia e caratteristiche individuali peculiari che implicano un indissolubile legame tra il testo stesso e il supporto su cui è scritto.

Si configura, quindi, un oggetto di studio complesso che richiede un'attenta analisi per coglierne gli aspetti rilevanti, e, dunque, la creazione di un'ontologia specifica che dia conto della complessità di questi materiali, laddove esistono invece modelli concettuali e ontologici già stabili per la descrizione dei testi a stampa, quali FRBRoo ([12]) e LRMoo ([20]).

Il modello che qui si presenta, CRMtex ([6]), fornisce tutti gli strumenti per gestire questo tipo di complessità definendo classi e proprietà per modellare tutte le operazioni coinvolte nella descrizione e nello studio di iscrizioni, papiri e manoscritti, dalla loro creazione nel passato, fino alla loro attuale conservazione, indagine e studio da parte degli studiosi, compresa la loro trascrizione, traduzione, interpretazione e pubblicazione.

Il modello è stato creato come estensione dell'ontologia CIDOC CRM ([2]), che si configura ad oggi come uno standard *de facto* nell'ambito dei Beni Culturali, soprattutto a livello europeo. Questo rende le informazioni modellate per mezzo di CRMtex perfettamente compatibili e del tutto integrabili con set di dati provenienti da diversi domini di ricerca e prodotte nell'ambito di varie iniziative internazionali.

La comunità scientifica coinvolta nello studio dei testi antichi, dai papirologi agli epigrafisti agli studiosi di manoscritti (antichi ma anche moderni ([11])), sta attraversando da tempo un intenso dibattito circa la necessità di dotarsi di modelli concettuali in grado di esprimere le entità complesse dei loro domini per mezzo di una codifica semanticamente ricca, al fine di stabilire l'interoperabilità dei loro dati con quelli generati in altri ambiti del sapere. Sebbene dal punto di vista semiotico (v. Il modello CRMtex) vi sia un unico meccanismo di produzione dei testi scritti, tradizionalmente lo studio dei testi antichi rientra in discipline diverse, generalmente cresciute intorno alle caratteristiche specifiche di ciascuna classe di documenti (es. papirologia per lo studio dei papiri ed epigrafia per le iscrizioni). Tuttavia, un approccio interdisciplinare e l'identificazione di elementi comuni sono fondamentali per conferire uniformità e

interoperabilità a tutte queste discipline, nonché per sfruttare competenze complementari provenienti da approcci diversi.

Lo sforzo di integrazione messo a punto da Papyri.info (175) e Trismegistos (175) e i vari tentativi compiuti da progetti come EAGLE (174) o iniziative quali Epigraphy.info (174) per sviluppare un modello semantico nel campo dell'epigrafia, testimoniano un interesse crescente per l'utilizzo di strumenti concettuali avanzati ed efficienti per la generazione di informazioni standardizzate, integrate e interoperabili in queste discipline ([15]). Nella stessa ottica, molte iniziative internazionali stanno focalizzando la loro attenzione anche sulle informazioni relative ai testi antichi e sulle sfide di interoperabilità in cui sono coinvolte. Iniziative europee quali PARTHENOS (175) e ARIADNEplus ([1]; [16]), stanno definendo e sviluppando archivi interoperabili basati sui principi FAIR ([9]) e profili applicativi in grado di mettere in relazione in maniera coerente i dati testuali con quelli archeologici, storici, artistici, ma anche con quelli scientifici e analitici, proiettandosi verso ambiti anche diversi da quelli dei Beni Culturali. Anche in questa prospettiva, lo sviluppo di sistemi in grado di interagire con questa mole eterogenea di informazioni diventa sempre più urgente.

Il CIDOC CRM e il suo ecosistema

L'attenzione alle problematiche della condivisione e dell'interoperabilità fra informazioni eterogenee ha animato lo sviluppo di CRMtex fin dalle sue origini. Uno dei punti di forza del modello consiste infatti nella sua piena compatibilità col modello CIDOC CRM, del quale si configura come estensione concettuale. Il CIDOC Conceptual Reference Model (CRM) è infatti un modello rilasciato dall'International Council of Museums (ICOM), certificato come standard ISO a partire dal 2014 (ISO21127:2014) e progettato per la modellazione di informazioni provenienti da musei, soprintendenze archeologiche, collezioni pubbliche e private, ministeri ed altri enti culturali simili. Attraverso due gruppi di lavoro, il Documentation Standards Working Group (DSWG) e il CIDOC CRM Special Interest Group (CIDOC CRM SIG), operativi già dagli anni '90, il CIDOC CRM è stato progressivamente arricchito e consolidato, divenendo una realtà scientifica riconosciuta a livello internazionale. Sin dai primordi l'obiettivo dei gruppi di lavoro è stato quello di creare un'ontologia formale event-oriented che potesse facilitare l'interscambio di informazioni eterogenee, come sono quelle del mondo della documentazione, in particolar modo in settori quali i Beni Culturali. Partendo da una visione di alto profilo, in cui il CIDOC CRM rappresenta il modello generale (*Core*), nel corso degli anni la ricerca ha portato all'elaborazione di numerose estensioni progettate per descrivere le entità di domini specifici ([3]), e andando a comporre quella che oggi è generalmente conosciuta come "CRM Family". Fra le estensioni più rilevanti si distinguono quella sviluppata nell'ambito della ricerca scientifica, CRMsci ([5]), che ha introdotto entità quali *S4 Observation*, mirata a rendere l'osservazione in senso scientifico di oggetti e fenomeni naturali, e *S10 Material Substantial*, orientata alla descrizione di elementi materiali del mondo fisico. Nell'ambito dell'archeologia, la realizzazione del modello CRMarchaeo ([4]) ha costituito una tappa fondamentale nello sviluppo dell'ecosistema CIDOC, fornendo uno strumento imprescindibile per la rappresentazione formalizzata di concetti archeologici troppo specifici per essere modellati per mezzo del *Core*. Entrami questi modelli sono stati presi come riferimento per lo sviluppo di CRMtex. Da

CRMsci, ad esempio, sono state ereditate le entità relative all'osservazione e all'analisi scientifica (la classe *TX5 Reading* di CRMtex è sottoclasse diretta di *S4 Observation* di CRMsci). CRMarchaeo ha invece fornito il materiale per la descrizione archeologica dei supporti materiali sui quali sono collocate le iscrizioni e i dati relativi al loro rinvenimento e alla loro classificazione.

Il modello CRMtex

CRMtex identifica e definisce in modo chiaro e univoco le principali entità coinvolte nello studio e nell'edizione di testi manoscritti antichi per poi descriverle in maniera formale mediante opportuni strumenti ontologici in una prospettiva multidisciplinare. La piena compatibilità di CRMtex con CIDOC CRM e le sue estensioni garantisce, infatti, l'interoperabilità persistente dei dati codificati per mezzo delle sue entità con altre informazioni semantiche prodotte dai Beni Culturali e dalle Digital Humanities.

Poiché la scrittura è un processo intellettuale finalizzato alla codifica di una lingua, CRMtex ha il suo fondamento scientifico negli aspetti semiotici del linguaggio e del testo ([10]). La modellazione, in particolare, si è basata sul funzionamento e sulle esigenze di rappresentazione delle scritture glottografiche (particolarmente fonografiche), per le quali sono state prese a campione scritture tipologicamente e semioticamente diverse relative a varie lingue indoeuropee. Il concetto centrale del modello è la nozione di 'testo' quale prodotto di un processo semiotico che implica un processo di codifica ('scrittura') e decodifica ('lettura'), attraverso una tecnologia umana particolarmente sofisticata, la scrittura, che consente la costruzione di un messaggio linguistico attraverso una serie di segni appositamente selezionati per questo scopo ([13];[14]). La scrittura appare, quindi, come un codice che richiede un processo di codifica da parte dello scrivente e uno di decodifica da parte del lettore per essere compreso correttamente. In questa prospettiva semiotica, un "testo" è costituito da una serie di segni tracciati fisicamente (cioè scritti) su un determinato supporto e destinati a codificare un'espressione linguistica con lo scopo dichiarato di comunicare uno specifico messaggio. Nella scrittura, come nella lingua, ogni componente (segno) è dotato di una duplice natura, fisica e concettuale: occorre, pertanto, distinguere tra la manifestazione fisica del testo, inteso come insieme di caratteristiche fisiche realizzate su un dato supporto attraverso l'uso di una tecnica di scrittura specifica (es. testo tracciato con l'inchiostro, dipinto, inciso, ecc.), dalla sua dimensione concettuale, cioè dall'insieme delle immagini mentali rappresentate da queste stesse caratteristiche fisiche, al quale ogni singola esecuzione deve essere ricondotta, sulla base di un principio di omogeneità, attraverso il quale riusciamo ad identificare un determinato segno indipendentemente dalla forma peculiare che il singolo scrivente gli fornisce.

I segni fisici che compongono il testo scritto costituiscono le manifestazioni materiali (glifi) delle unità del sistema di scrittura, ovvero i grafemi, le unità distintive funzionali minime della scrittura, le quali a loro volta codificano elementi linguistici di varia natura, in base al tipo di sistema di scrittura che una specifica lingua usa per notare se stessa. Ernst Pulgram affermava che «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» ([19]). Ad esempio: le singole lettere tracciate (glifi) dal lapicida che ha redatto l'iscrizione sull'Arco di Costantino rappresentano ciascuna un determinato grafema dell'alfabeto latino, ognuno dei quali a sua volta richiama – semplificando – un certo suono del sistema fonologico latino. In questi sistemi fonografici l'unità di base può essere diversa e il grafema, istanziato dal singolo grafo tracciato da un determinato scrivente, può fare riferimento ad altra unità linguistica, per esempio, una sillaba, come accade nel sillabario miceneo (la cosiddetta scrittura Lineare B) o nei sistemi cuneiformi in uso nel Medioriente antico.

Il processo di lettura si riferisce alla procedura semiotica della decodifica e quindi della comprensione di un testo scritto. Tale procedura può essere svolta a fini scientifici, allo scopo di analizzare e studiare il testo secondo diverse prospettive disciplinari. Dal lato del recupero del messaggio, poiché ogni grafema è legato a una data unità linguistica di lingue specifiche, la lettura del messaggio scritto presuppone la capacità di leggere la lingua di chi scrive.

Sul piano dei suoni linguistici, saranno i decodificatori (lettori, anche studiosi), che di volta in volta, sulla base della loro conoscenza del sistema linguistico, attribuiranno ad ogni segno o gruppo di segni l'adeguato valore linguistico. Nell'osservare un testo, quindi, è necessario tenere separata la procedura di decifrazione da quella di lettura. L'attività di lettura, quindi, è intesa come una specifica osservazione in cui si effettua la decodifica dei segni, cioè si riconosce il valore linguistico e si comprende il messaggio.

Nel progettare le entità di CRMtex abbiamo inizialmente indagato a fondo le interconnessioni esistenti tra il testo e le sue varie componenti. Sul versante del processo di lettura (cioè della decodifica del testo), e quindi dell'indagine del testo da parte degli studiosi, è rilevante l'individuazione anche di porzioni di testo, significative per le varie discipline, come, ad esempio, colonne, sezioni, paragrafi, singole parole o lettere o altri componenti specifici del testo scritto.

Gli studiosi delle diverse discipline, sulla base delle esigenze del loro studio, hanno, infatti, bisogno di selezionare e focalizzare la loro attenzione su diverse tipologie di porzioni testuali, per descriverne le condizioni fisiche (forma, disposizione, ecc.), per verificarne la leggibilità, o per rilevare particolari fenomeni (es. linguistici o paleografici) ad essi collegati. In un sistema ontologico è importante rappresentare in modo univoco la relazione logica e semantica della meronimia, collegando l'intero testo con le sue parti costitutive, ovvero le varie tipologie di porzione testuale individuate. In questo modo è possibile assegnare specifiche caratteristiche ai singoli segmenti di testo, indipendentemente dal testo nel suo insieme. Infatti, a singole porzioni possono essere associati particolari eventi di produzione o distruzione, come nel caso di lettere o parole danneggiate o usurate da agenti atmosferici o interventi umani.

CRMtex fornisce entità specifiche per descrivere tutti questi fenomeni, ed essendo fondata sul CIDOC CRM, sfrutta la potenza del suo ecosistema per descrivere e modellare informazioni generali, non testuali ma strettamente connesse ai testi e alla loro storia, come attori, luoghi, oggetti, eventi e le loro reciproche interrelazioni su base cronologica.

Il modello CRMtex si concentra anche sugli aspetti della ricerca scientifica e prevede classi e relazioni per descrivere le operazioni tipiche che studiosi di diverse discipline mettono in atto per acquisire conoscenze sui testi.

Oltre a prevedere una classe specifica per la lettura scientifica del testo (v. CRMtex: classi e relazioniIl modello CRMtex), il modello prevede una classe che permette di modellare i processi

di trascrizione del testo, operazione che si inserisce in un processo di interpretazione di segni scritti. La trascrizione può prevedere l'uso di un sistema di scrittura diverso da quello del testo originale, che si configura, quindi, come un ulteriore processo di codifica del testo. Questa operazione può essere accompagnata, inoltre, da un processo di traslitterazione, che implica una relazione univoca tra i segni dei due sistemi di scrittura. Tali processi sono coinvolti, per esempio, nello studio dei testi cuneiformi o, comunque, in generale, dei testi scritti in sistemi diversi dagli alfabeti latino e greco.

La ricca semantica che è possibile modellare risponde alla necessità di descrivere in dettaglio tutti gli eventi coinvolti nella vita del testo da codificare.

Per esempio, data la perdita di una porzione di testo, il modello, rispetto ad altri sistemi di codifica testuale, permette di specificare e collegare le circostanze storiche o ambientali che hanno determinato quella particolare condizione del testo, quando tali dati vengono recuperati in altri dataset o vengono alla luce durante il lavoro di ricerca; ad esempio l'evento di distruzione che ha prodotto la perdita potrebbe essere identificato come un evento storico documentato da altre fonti, la cui acquisizione permetterebbe a questo frammento di informazione di entrare a far parte di un più ampio grafo della conoscenza.

Integrazioni e correlazioni di questo tipo sono agevolate dal fatto che i dati modellati per mezzo di CRMtex (così come tutti quelli generati nell'ecosistema CIDOC CRM) sono agnostici rispetto a qualunque linguaggio di codifica formale. Le entità CRMtex possono infatti essere espresse secondo i linguaggi tipici del Web Semantico (quali RDF e OWL), codificate in formati di scambio comuni quali XML, Turtle o JSON (JSON-LD), essere distribuite e pubblicati come Linked Open Data. Questa flessibilità rende i dati testuali idonei per l'utilizzo in ambienti integrati e per l'impiego in qualunque scenario di interoperabilità basato sulla semantica. La flessibilità della codifica consente, inoltre, livelli più profondi di standardizzazione nella formalizzazione della conoscenza, per mezzo di arricchimenti successivi che possono essere implementati, ad esempio, attraverso l'uso di thesauri e vocabolari controllati, come quelli già ampiamente diffusi e utilizzati nell'ambito dei beni culturali.

CRMtex: classi e relazioni

Le classi e le proprietà di CRMtex individuano gli elementi ritenuti pertinenti alla modellazione e allo studio delle entità testuali, così come sopra definite. Al momento il modello presenta 9 classi e 11 proprietà, integrate nell'ecosistema ontologico del CIDOC CRM come sottoclassi e sotto-proprietà di elementi del *Core* o delle sue estensioni. Il modello, allo stato attuale, permette un'accurata descrizione dei sistemi di scrittura fonografici. Nelle descrizioni e negli schemi che seguono, le classi e le proprietà del CIDOC CRM *Core* sono indicati rispettivamente per mezzo delle lettere "E" e "P"; quelle dell'estensione CRMsci, con le lettere "S" e "O"; quelle del modello CRMarchaeo, con le lettere "A" e "AP". Per CRMtex sono state invece scelte le lettere "TX" per le classi e "TXP" per le proprietà, in modo da distinguere le specifiche entità del modello, pur mantenendo la piena compatibilità con l'ecosistema che lo ospita.

La classe *TX1 Written Text* (sottoclasse *E25 Man-Made Feature* del CIDOC CRM) descrive i segni fisicamente tracciati su un supporto allo scopo di comporre un testo. Tali segni sono modellati attraverso la classe *TX9 Glyph* (anch'essa sottoclasse di *E25*), che rappresenta la manifestazione concreta dei singoli segni tracciati dallo scrivente nell'atto di codificare un'espressione linguistica. I glifi sono tipicamente osservati dallo studioso durante un'attività di lettura (*TX5*, v. infra) svolta per decodificare e riconoscere i grafemi (*TX8*, v. infra) che rappresentano. L'appartenenza (fisica) di un segno (*TX9*) a un determinato testo scritto (*TX1*) è modellata attraverso la proprietà *TXP8 has component*, sotto-proprietà di *P46 is composed of*.

La classe *TX7 Segment* (sottoclasse di *TX1*) individua una porzione di testo di una qualsivoglia ampiezza ritenuta significativa dallo studioso per la propria indagine. Parti (*TX7*) e testo (*TX1*) sono collegati tra loro tramite la proprietà *TXP4 has segment*, sotto-proprietà di *P46 is composed of*.

La classe *TX2 Writing* (sottoclasse di *E12 Production*) descrive l'attività di creazione di segni su un supporto fisico attraverso l'uso di varie tecniche (pittura, incisione, ecc.) e di strumenti specifici (es. bulino, scalpello, penna, etc.). Il testo (*TX1*) e l'evento di produzione del testo (*TX2*) sono collegati attraverso la proprietà *TXP5 was written by* (sotto-proprietà di *P108 was produced by*).

La classe *TX3 Writing System* (sottoclasse di *E29 Design or Procedure*) rappresenta l'insieme convenzionale di segni e relative regole utilizzate per codificare e rappresentare (cioè scrivere) entità linguistiche allo scopo di convogliare un messaggio destinato ad essere recuperato a distanza di tempo e/o spazio da coloro che sono in possesso dello stesso codice (linguistico e scrittorio). Testo (*TX1*) e sistema di scrittura (*TX3*) sono collegati attraverso la proprietà *TXP9 is encoded by*, mentre il sistema di scrittura (*TX3*) e l'evento di scrittura (*TX2*) che ha prodotto il testo (*TX1*) sono legati tramite la proprietà *TXP1 used writing system*, sotto-proprietà di *P33 used specific technique*. Il sistema di scrittura (*TX3*) è, inoltre, collegato all'aspetto linguistico del testo tramite la proprietà *TXP6 encodes* (sotto-proprietà di *P2 has type*).

I segni del sistema di scrittura sono modellati attraverso *TX8 Grapheme* (sottoclasse di *E90 Symbolic Object*), che descrive le unità astratte con valore distintivo in un dato sistema di scrittura. L'appartenenza di un grafema (*TX8*) ad un dato sistema di scrittura (*TX3*) è modellato attraverso la proprietà *TXP7 has item*, sotto-proprietà di *P106 is composed of*.

La classe *TX4 Writing Field* (sottoclasse di *E25 Man-Made Feature*) rappresenta la porzione del supporto riservata ad accogliere il testo, talvolta fisicamente delimitata da elementi specifici, ad es. cornici o linee. La netta distinzione tra l'area contenente il testo scritto e le parti vuote del supporto (margini, *intercolumnia*, ecc.) e la tipologia di delimitazione dell'area è significativa per l'investigazione di elementi peculiari, fondamentali, ad esempio, per l'definizione degli stili e dei periodi di produzione dei documenti. La relazione tra *TX4 Writing Field* e il testo (*TX1*) è modellata attraverso la proprietà *TXP2 includes* (sotto-proprietà di *P56 bears feature*).

La classe *TX5 Reading*, come detto, è sottoclasse di *S4 Observation* di CRMsci e, come quest'ultima, è rivolta all'osservazione scientifica, la misurazione e la descrizione analitica tipica delle scienze descrittive ed empiriche. *TX5 Reading* descrive infatti la procedura semiotica di decodifica (e quindi comprensione) di un testo scritto, secondo criteri scientifici atti ad analizzare e studiare il testo secondo diverse prospettive disciplinari. L'attività di lettura è quindi intesa

come una specifica osservazione (*S4*) in cui si effettua la decodifica dei segni, cioè si riconosce il valore di ogni singolo segno e, di conseguenza, dell'intero messaggio. Testo (*TX1*) e lettura (*TX5*) sono collegati attraverso la proprietà *TXP10 was read by,* sotto-proprietà di *O8 observed* di CRMsci.

La classe *TX6 Transcription* (sottoclasse di *E7 Activity*), si riferisce all'attività di riscrittura del testo compiuta a fini scientifici da parte dell'editore. Questa operazione può comportare un sistema di scrittura (*TX3*) diverso da quello del testo originale, implicando una trasposizione dei segni da un sistema di scrittura all'altro (e, quindi, una ricodifica del testo). La trascrizione (*TX6*) del testo è collegata con l'attività di lettura (*TX5*) che ne è alla base attraverso la proprietà *TXP3 rendered*, sotto-proprietà di *P20 had specific purpose*. Attraverso la proprietà *TXP11 transcribed* (sotto-proprietà di *P16 used specific object*), viene modellato lo specifico modo in cui un'attività di trascrizione (*TX6*) rende i segni di un sistema di scrittura (*TX8*) istanziati in un testo (*TX9*).

Lo schema complessivo delle classi e delle proprietà di CRMtex è presentato in Figura 1 e Figura 2. In particolare, la Figura 1 presenta la modellazione del testo e la sua produzione, considerando i tre diversi livelli di codifica: il livello relativo al testo scritto, ovvero, i glifi (livello fisico); il livello relativo ai grafemi e alle altre entità simboliche codificate al momento della scrittura (livello simbolico); il livello relativo alle idee e ai concetti che il testo esprime (livello concettuale). La Figura 2 offre il punto di vista dell'indagine del testo con le entità legate alla sua lettura (cioè l'osservazione accurata delle sue caratteristiche fisiche) e alla sua trascrizione.

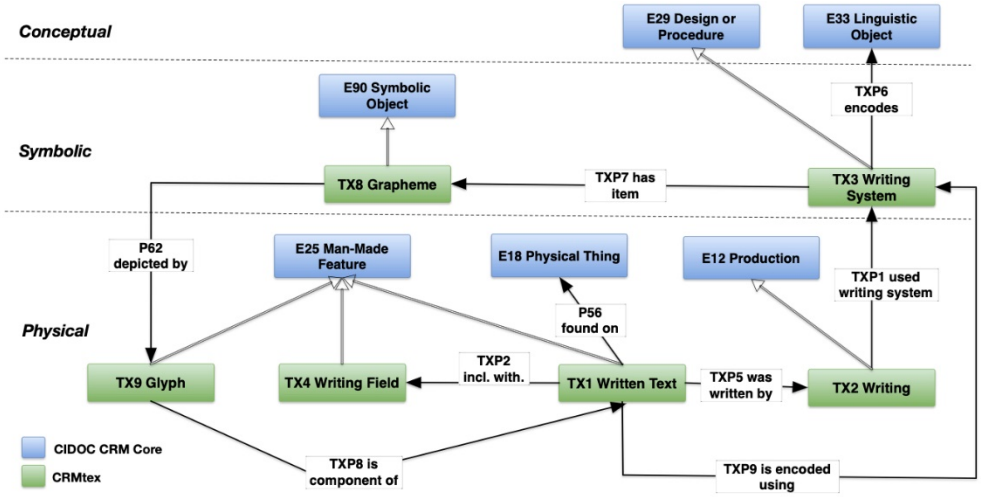

Figura 1: Eventi di scrittura e codifica del testo in CRMtex

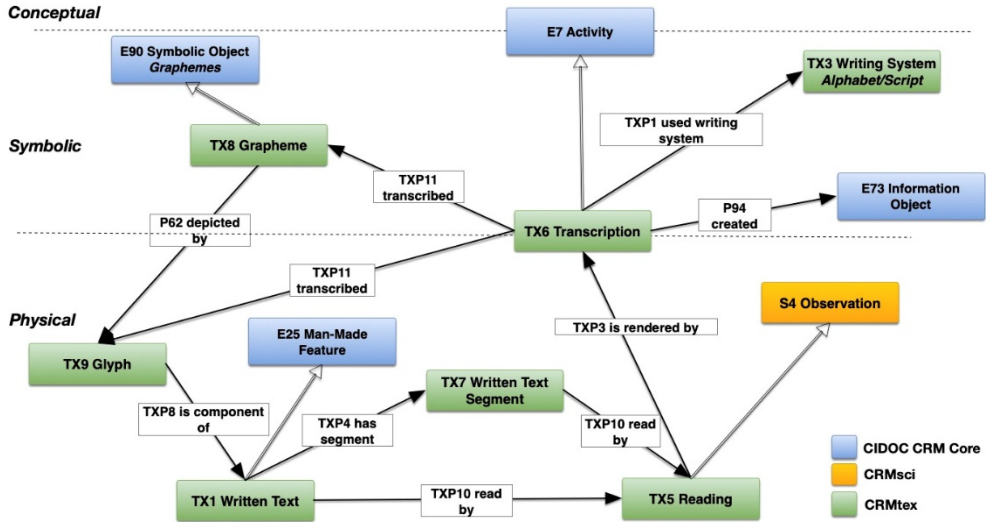

Figura 2: Lettura, decodifica e trascrizione del testo in CRMtex

Esempio di codifica CRMtex: la Tavoletta del Tripode

Per sostanziare quanto sopra descritto, si fornisce un esempio di modellazione delle informazioni pertinenti ad un testo epigrafico, la cd. "Tavoletta del Tripode" (PY 641). Si tratta di una tavoletta di argilla (26,5 x 3,8 cm) risalente al periodo del tardo bronzo (XIII sec. a.C.), scoperta nella stanza dell'archivio del palazzo miceneo di Pilo (odierna Englianos, Messenia) durante la campagna di scavo guidata da C.W. Blegen nel 1952. La tavoletta si trova attualmente conservata presso il Museo Archeologico Nazionale di Atene (num. inventario 709-712).

La tavoletta reca un'iscrizione su tre linee di testo in greco miceneo redatta, con andamento destrorso, in scrittura Lineare B. La scrittura Lineare B è una forma di sillabario (sillabogafia), un tipo di sistema fonografico in cui ogni segno (*TX8 Grapheme*) codifica una sillaba (*TX8 Grapheme -> P2 has type* → *E55 Type* → "*Syllabogram*"); a questi segni si aggiungono alcuni elementi logografici, come l'indicazione dei numeri, e alcuni elementi iconici.

Il testo consiste in una lista di oggetti di vasellame conservati nel palazzo di Pilo, al cui nome segue una rappresentazione iconografica del vaso e una indicazione numerica.

In Figura 3 si riporta il testo presente nella seconda parte della prima linea di iscrizione:

	testo della tavoletta (*TX1 Written Text* composto di *TX9 Glyph*)
	*TX8 Grapheme*
ti-ri-po e-me po-de o-wo-we *201VAS 1	traslitterazione del testo (*TX6 Transcription --> P2 has type --> E55Type*
tripode eme(i) pode(i) owowens *201VAS 1	trascrizione del testo (*TX6 Transcription --> P2 has type --> E55Type*
un tripode con un solo piede ansato [icona]	traduzione del testo (*P73 has translation --> E33 Linguistic Object*)

Figura 3: Frammento dell'iscrizione sulla Tavoletta del Tripode e suo studio

Le classi e le proprietà di CRMtex permettono la modellazione degli eventi di creazione del testo, i dati epigrafici, quelli relativi allo studio dell'iscrizione e del supporto in quanto oggetto archeologico. La rappresentazione schematica dell'interazione fra questi elementi è illustrata in Figura 4.

La figura mostra come il testo della tavoletta (*TX1 Written Text*) sia composto da una serie di segni fisici (*TX9 Glyph*), ognuno dei quali rimanda ad un'unità grafica astratta (*TX8 Grapheme*) che ne permette l'individuazione nonostante le peculiarità individuali che ogni scrivente pone nel tracciare i vari elementi grafici. L'appartenenza dei grafemi usati per redigere l'iscrizione al tipo di scrittura "Lineare B" è specificata per mezzo della classe *TX3 Writing System* e delle proprietà *TXP7 is item of* e *TXP1 used writing system*. Lo specifico evento di scrittura e le sue caratteristiche sono modellati per mezzo della classe *TX2 Writing* e della proprietà *TX5 was written by* che associa in modo univoco l'attività di scrittura con l'iscrizione da essa generata.

L'interpretazione dei segni avviene tramite un evento di lettura (*TX5 Reading*) in seguito al quale lo studioso provvede alla redazione di una traslitterazione e una trascrizione (*TX6 Transcription*) seguita da una traduzione del testo (*P73 has translation → E33 Linguistic Object*). Le proprietà *TXP10 read by*, *TXP3 is rendered by*, *P94 created* e *P14 carried out by* sono usate per dettagliare questa sequenza di eventi e definire gli attori (studiosi) coinvolti. La classe *TX3 Writing System* viene ulteriormente impiegata per indicare questa volta l'alfabeto latino, usato durante le varie fasi di studio del testo.

Le classi e proprietà del CIDOC CRM *Core* e delle altre estensioni, soprattutto quella archeologica (CRMarchaeo) e scientifica (CRMsci), consentono la descrizione di eventi correlati, quali il ritrovamento della tavoletta (*O9 was object found by → S19 Encounter Event*), lo scavo archeologico durante il quale tale ritrovamento è avvenuto (*A9 Archaeological Excavation*), gli attori coinvolti *(E39 Actor*). Consentono inoltre di modellare attività di ricerca specifiche, quali l'analisi delle componenti fisiche della tavoletta (composizione chimico-fisica, materiali,

provenienza) e del suo processo di produzione (*P108 was produced by → E12 Production*), oltre a quello delle tecniche di realizzazione e lavorazione, degli strumenti utilizzati ecc. L'uso delle entità temporali del CIDOC CRM (*P4 has time span → E52 Time Span*) consentono infine di distinguere e dettagliare i momenti chiave della storia della tavoletta, dalla sua produzione materiale e il suo utilizzo come supporto scrittorio, fino al momento del suo rinvenimento durante lo scavo del palazzo di Pilo (*S19 Encounter Event → AP25 occurs during → A9 Archaeological Excavation*).

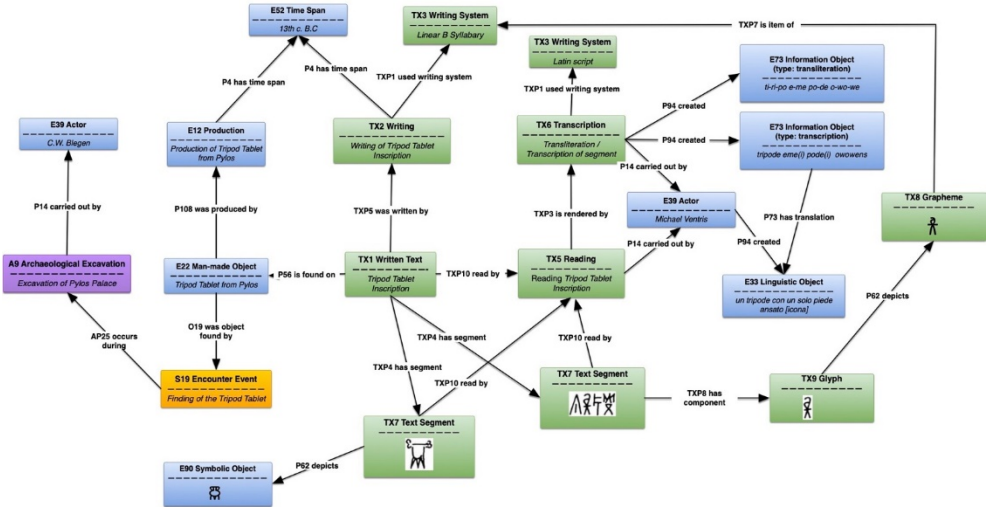

Figura 4: Modellazione dell'iscrizione sulla Tavoletta del Tripode in CRMtex

Sviluppi futuri

CRMtex è uno strumento in grado di conferire valore ontologico alle entità testuali, offrendo innumerevoli benefici per la ricerca in molte discipline umanistiche. La possibilità di fornire la rappresentazione dei dati culturali sul Web Semantico, di pubblicarli in formati standard (come LOD) e di renderli facilmente disponibili, interoperabili e riutilizzabili in un numero infinito di contesti, rappresenta sicuramente una delle sue caratteristiche più rilevanti.

Grazie anche ad esse CRMtex è stato scelto come modello per la codifica e l'integrazione di dati relativi a iscrizioni e graffiti nell'infrastruttura semantica che il progetto ARIADNEplus sta costruendo, ed è stato selezionato tra i modelli base per l'ontologia che l'iniziativa Epigraphy.info sta definendo per l'interoperabilità dei dati epigrafici.

Il modello è in continua espansione ed è orientato all'approfondimento di aspetti linguistici che possano valorizzarne ulteriormente le funzionalità e favorirne l'utilizzo in altre discipline. Il lavoro si sta attualmente focalizzando su una codifica semantica più dettagliata della relazione esistente tra grafema e unità linguistica codificata, tenendo conto non solo delle scritture fonografiche, ma delle scritture glottografiche nel loro insieme. In particolare, il lavoro si sta orientando verso la modellazione e la tipizzazione delle entità linguistiche notate attraverso la

scrittura. Un futuro sviluppo del modello consentirà di estendere la descrizione anche a sistemi di scrittura basati su modelli logografici e ideografici.

Un altro fronte operativo riguarda la definizione delle variazioni stilistiche e dei dati paleografici, che hanno grande importanza nella descrizione dei testi antichi, all'interno dei quali stili deversi possono venire tipicamente impiegati per scopi diversi e utilizzati in differenti tempi e luoghi, come avviene ad esempio per il corsivo tolemaico dell'Egitto ellenistico e la scrittura onciale maiuscola (III-VIII sec.). Questo aspetto è, quindi, fondamentale per la determinazione della datazione e dell'origine dei testi, soprattutto in riferimento ai singoli stili sviluppati in importanti centri di redazione e produzione di testi (ad esempio gli *scriptoria* dei monasteri) e necessita quindi di essere adeguatamente affrontato.

Al fine di potenziare l'integrazione e l'interoperabilità dei dati codificati con CRMtex, un ulteriore lavoro riguarda l'armonizzazione, già *in fieri*, con FRBRoo/LRMoo, modelli anch'essi compatibili con CIDOC CRM e volti a rappresentare la semantica alla base dell'informazione bibliografica. Molte classi di queste ontologie, infatti, trovano punti di contatto con CRMtex e potrebbero costituire la base per la creazione di uno strumento ontologico più complesso (ma più completo) per la modellazione efficace di entità testuali pertinenti sia a fonti manoscritte sia ad entità prodotte attraverso processi meccanizzati.

Acknowledgements

Il presente articolo è stato in parte realizzato con il supporto del Progetto PRIN 2017 "Lingue e culture dell'Italia antica. Linguistica storica e modelli digitali", finanziato dal Ministero dell'Università e della Ricerca.

References

[1] "ARIADNEplus Project." Accessed October 17, 2021. https://ariadne-infrastructure.eu/.
[2] "CIDOC Conceptual Reference Model." Accessed October 17, 2021. http://www.cidoc-crm.org/.
[3] "CIDOC CRM Compatible Models & Collaborations." Accessed October 17, 2021. http://www.cidoc-crm.org/collaborations.
[4] "CRMarchaeo. Excavation Model." Accessed October 17, 2021. http://www.cidoc-crm.org/crmarchaeo/ .
[5] "CRMsci. Scientific Observation Model." Accessed October 17, 2021. http://www.cidoc-crm.org/crmsci/ .
[6] "CRMtex. Model for the Study of Ancient Texts." Accessed October 17, 2021. http://www.cidoc-crm.org/crmtex/ .
[7] "EAGLE. Europeana Eagle Portal." Accessed October 17, 2021. https://www.eagle-network.eu/.
[8] "Epigraphy.Info." Accessed October 17, 2021. https://epigraphy.info/.

[9] "FAIR Principles." Accessed October 17, 2021. https://www.go-fair.org/fair-principles/.
[10] Felicetti, Achille, and Francesca Murano. 2021. "Ce Qui Est Écrit et Ce Qui Est Parlé. CRMtex for Modelling Textual Entities on the Semantic Web." *Semantic Web* 12, 2: 169–80. https://doi.org/10.3233/SW-200418.
[11] Felicetti, Achille, and Francesca Murano. 2016. "CRMtex: Semantic and Semiotic Strategies for the Encoding of Hand-Written Documents." In *Atelier "Les Manuscrits de Saussure, Parmi d'autres."* Genève.
[12] "FRBRoo. Functional Requirements for Bibliographic Records." Accessed October 17, 2021. http://www.cidoc-crm.org/frbroo/home-0.
[13] Harris, Roy. 1993. *La Sémiologie de l'écriture. CNRS Langage*. Paris: CNRS éditions.
[14] Harris, Roy. 2013. *Signs, Language and Communication*. London: Routledge.
[15] Liuzzo, Pietro Maria, and Silvia Evangelisti. 2021. "Modeling Execution Techniques of Inscriptions." *Semantic Web* 12, 2: 181–90. https://doi.org/10.3233/SW-200395.
[16] Meghini, Carlo, Roberto Scopigno, Julian Richards et al. 2017. "ARIADNE: A Research Infrastructure for Archaeology." *Journal on Computing and Cultural Heritage* 10, 3: 1–27. https://doi.org/10.1145/3064527.
[17] "Papyri.Info." Accessed October 17, 2021. https://papyri.info/.
[18] "PARTHENOS Project – Pooling Activities, Resources and Tools for Heritage E-Research Networking, Optimization and Synergies." Accessed October 17, 2021. http://www.parthenos-project.eu/.
[19] Pulgram, Ernst. 1976. "The Typologies of Writing-Systems." In *Writing Without Letters*, edited by William Haas: 1–28. Manchester-Totowa (NJ): Manchester University Press-Rowman and Littlefield.
[20] Riva, Pat, and Maja Žumer. "FRBRoo, the IFLA Library Reference Model, and Now LRMoo : A Circle of Development." In Transform Libraries, Transform Societies. Kuala Lumpur, Malaysia: IFLA Library, 2018. http://library.ifla.org/id/eprint/2130/1/074-riva-en.pdf.
[21] "Trismegistos. An Interdisciplinary Portal of the Ancient World." Accessed October 17, 2021. https://www.trismegistos.org/index.php.