

ARCHEOLOGIA E CALCOLATORI

29

2018

All'Insegna del Giglio

ARCHEOLOGIA E CALCOLATORI

29

2018

All'Insegna del Giglio

Realizzazione grafica della sovracoperta di Marcello Bellisario
Rivista «Archeologia e Calcolatori» (ISSN 1120-6861, e-ISSN 2385-1953)
ISBN 978-88-7814-795-9, e-ISBN 978-88-7814-796-6
© 2018 – All’Insegna del Giglio s.a.s. – www.insegnadelgiglio.it
Firenze, novembre 2018
Stampa, Andersen S.p.a.
Abbonamento: € 40,00. Spedizione: Italia, gratuita; estero, a carico del destinatario.

I dati forniti dai sottoscrittori degli abbonamenti vengono utilizzati esclusivamente per l’invio della pubblicazione e non vengono ceduti a terzi per alcun motivo.

INDICE

I DATI ARCHEOLOGICI. ACCESSIBILITÀ, PROPRIETÀ, DISSEMINAZIONE
(Roma, CNR, 23 maggio 2017), a cura di MARCO ARIZZA,
VALERIA BOI, ALESSANDRA CARAVALE, AUGUSTO PALOMBINI,
ALESSANDRA PIERGROSSI

INTRODUZIONE	9
LUIGI MALNATI, <i>Il meglio è nemico del bene</i>	13
ALBERTO MARIA GAMBINO, MARIA LETIZIA BIXIO, <i>I dati archeologici tra prerogative di tutela e istanze di accesso</i>	19
MIRELLA SERLORENZI, <i>Accessibilità e diffusione del dato archeologico: l'esperienza del SITAR</i>	31
VALERIA BOI, <i>Archeologia professionale fra diritto d'autore e accesso ai dati</i>	41
GABRIELE GATTIGLIA, <i>La maternità e la paternità del dato archeologico. L'esempio del MOD (Mappa Open Data)</i>	51
MARCO CIURCINA, <i>Riflessioni sul diritto d'accesso ai beni culturali</i>	59
ELENA CALANDRA, VALERIA BOI, <i>Tra riproduzione e condivisione dei beni culturali: il ruolo dell'Istituto Centrale per l'Archeologia</i>	63
MIRCO MODOLO, <i>Promozione del pubblico dominio e riuso dell'immagine del bene culturale</i>	73
AUGUSTO PALOMBINI, <i>Riproducibilità a vario titolo del patrimonio: situazione e prospettive</i>	87
PAOLA MOSCATI, <i>La diffusione della cultura scientifica: documentazione e disseminazione nei progetti di ricerca dell'ISMA</i>	93
MARCO ARIZZA, ALESSANDRA CARAVALE, ALESSANDRA PIERGROSSI, <i>Accessibilità e disseminazione dei dati nell'esperienza dell'ISMA</i>	101
SILVIA CHIODI, <i>Aspetti etici nel campo dell'open access</i>	111

*

* *

FRANCESCO DI FILIPPO, MASSIMO MAIOCCHI, LUCIO MILANO, RENZO ORSINI, <i>The "Ebla Digital Archives" Project: how to deal with methodological and operational issues in the development of cuneiform texts repositories</i>	117
MARIANNA FIGUERA, <i>Database management e dati archeologici: standardizzazione e applicazione della logica fuzzy alla gestione delle fonti e delle attribuzioni tipologiche</i>	143

MIGUEL BUSTO-ZAPICO, ENRICO CIRELLI, <i>Indagine sul fenomeno della standardizzazione nelle produzioni ceramiche provinciali in epoca tardo romana. La forma Hayes 85</i>	161
GERVASIO ILLIANO, Apud Baulos. <i>L'utilizzo della viewshed analysis per la risoluzione di un quesito di topografia flegrea</i>	185
NATASCIA PIZZANO, PASQUALE MEROLA, ELISA DI GIOVANNI, <i>Archeologia preventiva a Pontelatone (CE): nuovi dati sulle dinamiche insediative in prossimità del fiume Volturno</i>	201
ALBERTO BELUSSI, SARA MIGLIORINI, PIERGIOVANNA GROSSI, <i>The Archaeological Urban Information System of Verona: an approach to interoperability through standard-based conceptual modelling</i>	223
ANDREA FIORINI, <i>Il metodo fotografico RTI (Reflectance Transformation Imaging) per la documentazione delle superfici archeologiche. L'applicazione ai materiali di età protostorica</i>	241
PAOLO RONDINI, <i>Digital Rocks. An integrated approach to rock art recording: the case study of Ossimo-Pat (Valle Camonica), monolith 23</i>	259
MARTINA GIANNINI, CRISTINA CASTAGNETTI, <i>Un approccio multidisciplinare per documentare e visualizzare un sito archeologico: la villa d'otium di Massaciuccoli (Massarosa, LU)</i>	279
MANUELA INCERTI, SARA D'AMICO, STEFANO GIANNETTI, GAIA LAVORATTI, ULIVA VELO, <i>Le Digital Humanities per lo studio e la comunicazione di beni culturali architettonici: il caso dei mausolei di Teodorico e Galla Placidia in Ravenna</i>	297
ALBERTO URZIA, SIMONE ZAMBRUNO, ANTONINO VAZZANA, MICHAEL ANDERSON, COLLEEN M. DARNELL, <i>Prototyping an Egyptian revival. Laser scanning, 3D prints and sculpture to support the Echoes of Egypt exhibition</i>	317

Recensioni:

S. ORLANDI, R. SANTUCCI, F. MAMBRINI, P.M. LIUZZO (eds.), *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference*, Roma 2017, Sapienza Università Editrice (I. Rossi), p. 333; V. BIGOT JULOUX, A.R. GANSELL, A. DI LUDOVICO (eds.), *CyberResearch on the Ancient Near East and Neighboring Regions. Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, Leiden 2018, Brill (I. Rossi), p. 337; C. DAL MASO (ed.), *Racconti da museo. Storytelling d'autore per il museo 4.0*, Bari 2018, Edipuglia – A. FALCONE, A. D'EREDITÀ (eds.), *Archeosocial. L'archeologia riscrive il web: esperienze, strategie e buone pratiche*, Mozzecane (VR) 2018, Dielle Editore – S. PALLECCHI, *Raccontare l'archeologia. Strategie e tecniche per la comunicazione dei risultati delle ricerche archeologiche*, Firenze 2017, All'Insegna del Giglio (I. Pompili), p. 342.

THE “EBLA DIGITAL ARCHIVES” PROJECT:
HOW TO DEAL WITH METHODOLOGICAL
AND OPERATIONAL ISSUES IN THE DEVELOPMENT
OF CUNEIFORM TEXTS REPOSITORIES

1. INTRODUCTION¹

Within organized collections of documents in the ancient world, the Ebla evidence stands out as a privileged case study. In fact, its great antiquity – 24th century BCE – makes it the first known archive in the history of mankind, for which extensive archaeological information on its primary setting is available (WELLISH 1981; ARCHI 2003). In addition, the number of texts retrieved is very conspicuous, roughly reaching 3,500 tablets². As an added benefit, the texts belong to various genres, which often overlap (ARCHI 1986): administrative texts, legal documents, “letters”, bilingual vocabularies, literary compositions, ritual texts, etc. The historical significance of the Ebla archives can hardly be overrated. The astonishing amount of information provided by the texts is however of difficult access, due to the inherent difficulties offered by both language and writing system (KREBERNIK 1996; HUEHNERGARD, WOODS 2004; RUBIO 2006; CATAGNOTI 2012). The aim of this article is to offer an overview of the specificities of the Ebla sources and how they impact in the development of digital tools for the analysis of ancient texts. A few methodological issues will be raised in the second part of the article, where a description of the Ebla Digital Archives project (EbDA) is provided in some detail.

2. THE WRITING SYSTEM OF THE EBLA TEXTS

In order to better appreciate the EbDA database structure, as well as the scripting techniques used to populate and manage the records, a cursory description of the writing system is in order. The Ebla documents are written in logo-syllabic cuneiform, which is not alphabetic in nature. This sophisticated writing system requires complex data handling and *ad hoc* solutions, aimed to capture all of its features. The Ebla texts are to some extent close to archaic cuneiform from the late 4th-early 3rd millennium BCE, in that they usually provide very limited grammatical information. For instance, verbal

¹ In a spirit of collaborative work, M. Maiocchi and L. Milano prepared §1-§2.7; F. Di Filippo §3-§3.4; R. Orsini §4-§4.4. The conclusions of §5 are the shared product of all authors. The Ebla Digital Archives Project is freely accessible online at <http://ebda.cnr.it>.

² This is an estimate of the total number of complete tablets. The total number of inscribed objects (fragments and complete tablets) reaches roughly 12,000 items.

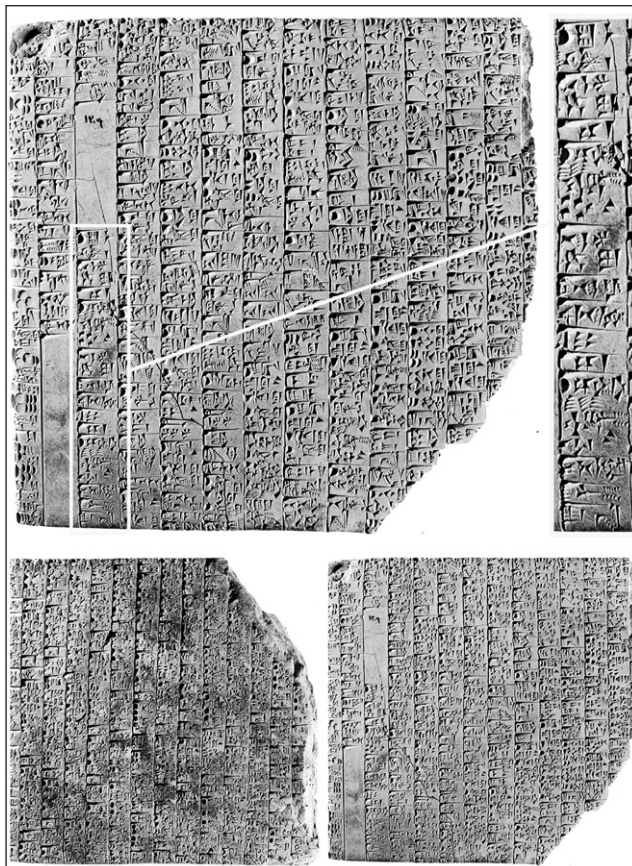


Fig. 1 – A cuneiform text from Ebla (= ARET I 1). Bottom: obverse (left) and reverse (right) views; top: magnified view of reverse (left) and detail of column XII (right).

forms may appear as a sign representing the bare root (i.e. a logogram, defined below), regardless of the actual tense or verbal aspect that is required in a given context: for instance, the sign DU, which originally is the pictographic representation of a human foot, may be used to express either “I go”, “you go”, “he goes”, etc. or “I went”, “you went”, “he went”, etc. Text layout at Ebla is also archaic, especially with regards to administrative texts, which form the most of the archive. Information is usually arranged in columns, each containing several boxes (i.e. lines of uneven size; Fig. 1).

Each box usually contains a semantic whole, such as for instance a number plus a noun to which it refers, a verbal form, a preposition, etc. Contrary

to this practice, information in later cuneiform texts is arranged in lines of even size. Punctuation is absent, although blank spaces of variable length are sometimes used to divide the text in logical units (including subtotals, totals and colophons). Logo-syllabic cuneiform (see definitions below) records information by means of a rather large set of visual symbols – usually ranging from a couple of hundred items up to a couple of thousand, depending on time, region, and text *corpora*. Each sign, considered as an independent entity without any context, is referred to by a sign-name, which is derived by one of its commonly attested values (either semantic or phonetic, see below). Sign-names are conventionally represented in upper-case characters (e.g. the sign for “day” is UD). More in detail, signs may be grouped into classes depending on their function, as described in the following paragraphs.

2.1 *Logograms: writing words according to their meaning*

Signs used to express words are commonly referred to by philologists as logograms, or word-signs, although a better definition would be morphograms (a morpheme is defined as the smallest part of a word carrying meaning). The earliest cuneiform documents from Southern Mesopotamia were originally logographic in nature, with very few exceptions. For instance, to write down the word for day, Ebla scribes may use the UD sign – originally a pictographic representation of the sun rising from the mountains – probably read /yawmum/ in Eblaic. Similarly, the word for night may be written with a compound logogram made of two signs, namely MI and AN: the first element, i.e. the sign MI, is a pictographic representation of a rainy cloud, conveying the meaning dark; the second element AN is a pictographic representation of a star, conveying the meaning sky. Thus, the semantic compound dark+sky is used to express night, and is read /mūsum/ in Eblaic. Semantic analysis of other compound logograms proves however often difficult. Depending on their origin, logograms can be further divided in Sumerograms, Akkadograms, and Eblaitograms (see below §2.7).

2.2 *Syllabograms: writing words according to their sound*

Signs standing for syllables are referred to as syllabograms. By definition, syllabograms convey phonetic information only. For instance, the word for day, besides as a logogram, is attested in the syllabic spelling *a-wa-mu*, standing for /yawmū/ “days” (on this spelling, see comments below). The word for night is also spelled *mu-šum* for /mūsum/, etc. In order to keep the total amount of signs within a manageable range, not all possible syllables and/or phonetic clusters are represented by a dedicated sign. Therefore, scribes have to make decisions on how to represent “problematic” words, such as the above mentioned *a-wa-mu*. In such spelling, the first two signs express one syllable (i.e. /^{*}yaw-/), and the syllabogram WA = *wa* is exploited here for its consonantal part only. As a rule of thumb, at Ebla vowel signs (such as A)

may be used to express syllables of the type /HV/, /VH/ and /HVH/, where V = vowel, and H is a weak consonant (KREBERNIK 1982, 179; CATAGNOTI 2012, 8-10). For instance, the word for “where?” is spelled *a* for /ʾay/. The same spelling may represent the word for “or” as well, i.e. /ʾaw/ – readers must differentiate the meaning on the basis of context. As phoneticism is often only loosely represented in writing, philologists must use great care in the reconstruction of the language.

2.2.1 Phonetic complementation and disambiguation

As most logograms are usually attached to more than one meaning (and therefore reading), sometimes syllabograms are used in conjunction with logograms to help the reader decide what logographic value is meant in a given context. For instance, the sign TUG₂ – originally the pictographic representation of a loom – stands for either tug₃ /tug/ “textile”, or mu₄ /mu/ (also read mur₁₀ /mur/) “to dress”. In order to disambiguate, Ebla scribes usually write down the latter word as either TUG₂.MU or MU.TUG₂, where MU stands for the syllable /mu/. As a convention, syllabic signs used in such way are represented in superscript in modern transliterations, respectively mu₄^{mu} and ^{mu}mu₄ in the example above. Such typographic convention is unfortunate, as it causes ambiguity with the representation of determinatives, also in superscript (see the following paragraph). However, phonetic complements and determinatives are encoded differently, therefore preserving the information on the function of these signs. Both deserve in their own right special attention.

2.3 Determinatives: semantic complementation

Further help in reducing the inherent ambiguity of possible readings attached to cuneiform inscriptions is provided by determinatives. Determinatives are defined as signs that are not meant to be read aloud, but help the reader decide the meaning of a sign or sequence of signs, occurring either in front or after the determinative associated to them. Determinatives carry a function similar to phonetic complements, but they work on the semantic domain. For instance, divine names are often preceded by the AN sign, in which case it is conventionally represented by superscript “d” (for Latin *deus* = god) in transliteration. Accordingly, the name of the storm god Hadda is transliterated ^da₃-*da* (see below for the use of the subscript number). Similarly, wooden objects are preceded (and sometimes followed) by the GĪŠ sign, which originally represents a wooden log, etc. In some instances, determinatives also help the reader decide the reading of a sign or sequence of signs. Going back to the spelling of the name of the storm god at Ebla, ^da₃-*da* corresponds to the sign sequence AN.E₂.DA. In turn, the sign E₂, when taken in isolation, is also used as a logogram for house, read /baytum/ in Eblaic. However, the presence of the determinative for divine names, i.e. the star sign AN, gives the

reader a hint to the fact that what follows is a divine name, thus excluding a reading somehow connected with the word for house.

2.4 Polysemy, polyphony and homophony

As it appears from the discussion above, usually a given sign is associated with multiple logographic values, and multiple syllabic values as well. This way, scribes could reduce the total amount of graphemes to be memorized. For instance, the KA sign, originally a pictographic representation of a human head with extra stippling marking the mouth, may be used to represent the words for “mouth”, “tooth”, “word”, etc. This property of signs is called polysemy. Similarly, the sign GA at Ebla is used to represent the syllables /ga/, /ka/, /qa/, /ġa/ (potentially also /gaH/, /kaH/, /qaH/, and /ġaH/ as well, see above §2.2) etc., according to the so-called polyphony principle. Polysemy and polyphony in turn fall under the comprehensive umbrella of the polyvalence principle, stating that a given sign may be associated with more than one function (i.e. it may occur as a logogram, syllabogram, or determinative). For instance, the star sign, besides being used as a determinative for divine names, may also stand for the word for “god” (*ʾilum/* in Eblaic), or for the syllabic value /an/. Similarly, the KA sign, may be used for the syllabic value /bu/, etc.

Conversely, two or more distinct signs may end up having the same (or very similar) reading, in which case they are conventionally distinguished in modern transliterations by a lowercase numerical index, and/or accents corresponding to subscripts 2 and 3, as illustrated in Table 1. As additional examples, at Ebla the syllable /bu/ may be expressed either with the sign KA (= *bu*₁₄), or BU₃ (= KA × “ŠU” = *bu*₃), or NI (= *bu*₁₆), or MUNU₄ (= *bu*_x), and in some special cases with BU = *bu* as well. This phenomenon, called sign homophony, is due to the complex history of the development of language and writing, which cannot be followed here for reasons of space.

2.5 Notes on graphemics and allography

Graphemes, i.e. distinct minimal units within the sign *corpus*, may be arranged in a number of different ways: inclusions (partial or total), juxtaposition, ligature, crossing, etc. For instance, the sign KU₂, which is used to express the verb “to eat”, is composed by two graphemes: the sign for “mouth” and the sign for “food” (originally a pictographic representation of a vessel for rations), placed either within the former, or in close proximity to it. However, in the latter case an interpretation in terms of a different compound logogram, namely KA.GAR = *inim gar* “(to make a) legal claim” (lit. to place/put a word), is also possible. This may be an extreme case, but uncertainties in the interpretation of the documents may suggest to leave the readings of some sign or sign sequence open, i.e. using sign names instead of possible




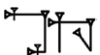

Sign	Sign name	Syllabic reading	Logographic reading	Determinative
	Ú = U ₂	ú for / ^h u/, / ^h u/, /lu/		ú “grass, plant”
	Û = U ₃	ù for / ^h u/, / ^h u/, also / [?] u/, rarely /wu/		
	U ₄	u ₄ for /yu/	u ₄ “day”	
	MA ₂ .HU (U ₅)	-	u ₅ “(a kind of courier)”	
	EZENxBAD (U ₉)	u ₉ for /hu/, /hu/, /yu/, rarely also /lu/	bàd “wall”	

Table 1 – Sign homophony. Drawings of cuneiform signs courtesy of Erica Scarpa.

values attached to the individual signs. Similarly, the sign DIRI is composed of two parts, namely SI and A, which can be arranged either as A.SI or SI.A. In both cases reading and meaning remain unchanged: DIRI = diri, which at Ebla almost invariably means “(to be) abundant, surplus”. Distinct graphic forms taken by a given sign are labelled allographs³.

2.6 Ancient spellings and modern transliterations

As a convention, the two possible forms of DIRI are transliterated as diri(A.SI) and diri(SI.A) respectively. More in general, parenthesis may be used to provide information on what is actually observable on the tablet. For example, the notation di[ri](SI.[A]) clarifies that the final part of the sign is broken – square brackets mark missing text. As writing is a human product, it is subject to variation, complications, and errors. Unexpected spellings occur quite often within the Ebla *corpus*. In transliteration, these are marked by an exclamation mark. For instance, the word for “weaver” is invariably written tug₂-nu-tag, except for a few instances having tug₂!(KU)-nu-tag, i.e. the scribe mistakenly wrote KU instead of TUG₂, the two signs being graphically very similar. Careful notation of what is actually written on the tablet not only preserves information, but opens up the possibility of searching for patterns in aberrant spellings that may have implications in terms of relative dating of

³ Allographs are not limited to ancient logographic scripts. In most alphabetic scripts, uppercase and lowercase instances of a given character are in fact allographs.

the tablets, scribal conventions, and palaeography. Interesting cases from the point of view of Assyriology include: 1) the personal name *dag-mul-da-mu*, written also *dag!(DAG×KASKAL)-mul-da-mu*, *dag-mul!(AN.MUL)-da-mu*, and *dag!(DAG×PAP)-mul!(AN.MUL))-da-mu*; 2) the sign U₃ written IGI.ŠÈ (as opposed to IGI.DIB) in *ù!(IGI.ŠÈ)-ma*; 3) the remarkable frequency of the use of ME instead of IGI in *igi!(ME)-sig*; 4) the spelling *lu-ma-na-du-ma!(NA)* where the mistake is apparently phonetic in nature.

This complicated situation obviously impacts on the searchability of the archive content (see however below §4.3). In this regard, legacy transliterations being used by scholars for reasons of clarity are also problematic, for a twofold reason: 1) the input of texts having outdated readings may potentially introduce incoherent values (despite of all scholarly efforts, as the amount of signs and spellings to control is very large); 2) as the reading of a given sign often depends on personal preference and expertise of the individual scholars, queries within the database may potentially not yield the expected results, as the user may not know in advance what reading is used within the database. For instance, the word for “gold” is written *ku₃-sig₁₇*, but it is commonly transliterated *ku₃-gi* in several scholarly editions of Ebla material (*sig₁₇* and *gi* are both possible values attached to the sign GI). In order to mitigate this situation, we developed queries based on sign names (see below §3.3 and §4.4): the query input (either *ku₃-sig₁₇* or *ku₃-gi* in the example above) is transformed in the corresponding sign sequence (KU₃ GI), and the matching results are returned. This technique is also useful to query text having unclear interpretation.

2.7 *Alloglottographic systems: stratified writing traditions*

Cuneiform writing system was invented almost a millennium earlier than the first Ebla texts, probably by Sumerians in modern-day southern Iraq. Sumerian is an isolate language, having a different grammar, as well as a different set of contrastive sounds (phonemes) compared to Eblaic and more in general to Semitic languages. Therefore, Semitic scribes had to adapt the writing system to write down their own language. A word written down according to its original Sumerian spelling, but read in the local language, is also labelled a Sumerogram (all Sumerograms are logograms). Ebla scribes however didn't get their writing system directly from the Sumerians, but from the Akkadians⁴, another Semitic population probably originally stemming from somewhere in the eastern Tigris area.

Together with the writing system developed by the Sumerians, Ebla scribes took the habit of writing down a few words according to their (frozen)

⁴ More precisely, Ebla got its script probably from Mari, SALLABERGER 2001.

Akkadian spelling. Such words are labelled Akkadograms, and are usually transliterated in uppercase italic. For instance, the word for “man, people” is written *NA-SE₁₁* at Ebla, corresponding to the local /nisay(n)/⁵. Finally, at Ebla a few words appear to be written down in frozen forms that however reflect the local lexicon. Such words are labelled Eblaitogram, transliterated here in uppercase italic underlined. For instance, the queen of Ebla is referred to as *MA-LIK-TUM*, probably read /malkatum/⁶. The presence of Sumerograms, Akkadograms, and Eblaitograms makes the writing system of Ebla an alloglottographic one. In such system, words may be written down according to their spellings in foreign languages (RUBIO 2007).

3. DESIGNING THE ARCHITECTURE OF A DIGITAL REPOSITORY FOR CUNEIFORM TEXTS

The design and architecture of a digital repository of richly-annotated text editions *per se* is a challenge in many respects. The construction of such a digital resource for cuneiform texts in transliteration⁷ must face several additional issues (DEL FREO, DI FILIPPO 2014; DI FILIPPO 2018). The necessity of preserving all information levels within a logo-syllabic writing system, the fact that it only loosely represents the actual spoken form of the underlying language, the necessity of a multi-level architecture to manage annotations on different textual entities (from an entire document to individual cuneiform signs, see below §3.4): these facts have forced us toward the development of an *ad hoc* solution for complex data handling, aimed at capturing the richness of the Eblaite *corpus* as a whole. In particular, the project has entailed the prototyping of a solution for digitization of texts, including parsing of shallow annotated text files passed as input, setting up of a suitable data repository (including both text and photos of the inscribed objects), and the realisation of a graphical user interface for inserting, querying and annotating structured data.

In doing that, we identified four analytical levels for the representation of the individual documents and of the archive as a whole: a tablet-level, a word-level, a sign-level, and an annotation-level.

⁵ The attested Eblaic form is grammatically a dual. The presence of SE₁₁ in the spelling of NA-SE₁₁ is also an indication that this is not a local syllabic spelling. SE₁₁ is probably an Akkadian development that does not conform to the local scribal tradition (CATAGNOTI 2012, 7). In addition, NA-SE₁₁ occurs also in reduplicated form to express plurality, i.e. it is treated as a logogram (plurality in Eblaic is otherwise expressed appending the appropriate case ending written syllabically).

⁶ This word cannot be an Akkadogram, as the word for “queen” in Akkadian is *šarratum*.

⁷ An effective computational approach in the field of cuneiform epigraphy is yet to come, especially regarding optical character recognition (OCR).

3.1 *Tablet-level*

In order to fully unveil the potential of the Ebla archives in terms of historical research, the documents must be reorganized as much as possible into their original scopes. This is however not straightforward in Ebla studies, as the individual fragments were often edited in different overlapping series (i.e. MEE and ARET), and sometimes joined together and re-published as new texts (e.g. the volume ARET 12).

With these considerations in mind, our first concern focused on the development of a serviceable querying tool combining archaeological data (i.e. context, chronology, and findspots) and typological information, to be further related with data pertaining to deeper levels of information (word-level and sign-level). Each document is thus stored with comprehensive metadata annotations and, in addition, bibliographic references. This constitutes the backbone catalogue of our database, which is however under constant update, in order to take track of possible new editions and textual joins of previously published material.

3.2 *Word-level*

As a further step in the digital processing of cuneiform texts from Ebla, transliterated texts must be parsed into meaningful units. Thus, a tablet may be segmented into paragraphs and subparagraphs (however defined), (sub) paragraphs into words, words into signs, signs into graphemic constituents. It is worth noting, however, that high order text mining techniques are presently unsatisfactory when applied to cuneiform material, due to the absence of controlled vocabularies, and dedicated softwares such as lemmatizers and chunkers⁸. We therefore opted for an approach based on content tokenization, consisting in the segmentation of text into its basic semantic components, broadly defined as to include meta-textual information (such as for instance indication of broken textual parts). Such process produces a digital representation of a given document in terms of an exhaustive and unabridged set of its word-level features, alongside their positional information, which in turn represent the starting point for more refined text processing and information retrieval methodologies.

The identification of these minimal semantic units (tokens) is in fact a challenging process, because of the complexity of the cuneiform writing system, as well as to the way it is rendered in modern transliterations (§2). Even for modern languages and writing systems, in fact, the main challenge

⁸ So far, only few attempts of high-level text processing have been attempted (e.g. MACKS 2002): all of them, however, came up against the manifold issues of the cuneiform logo-syllabic script, but above all they cannot be easily integrated into a wide scale text processing as the case of Ebla archives.

in identifying token boundaries involves the ability to recognize meaningful patterns, rather than simply relying on whether a sequence of characters is bounded by delimiters on either side.

Let us consider for instance the following small portion of the text ARET 3, 42 (r?.2'.0-r?.2.2):

1. (the beginning of column is broken)
2. [•] 'a₃-da-um^{tug2}-II
3. 1 gu-zi-tum^{tug2} 1 ib₂-III-dar-sa₆^{tug2}

Line 1 does not preserve any word-level feature, but merely a notation referring to the state of preservation of the tablet. Therefore, it would be inconsistent to consider here whitespace characters as token boundaries. As for Line 2, the first unit is transliterated with four glyphs: “[”, “•”, whitespace, and “]”, standing for a broken part of text maintains one cuneiform sign. This unit must be considered as a token, as it preserves the positional information of the following word ('a₃-da-um^{tug2}-II), preserving at the same time the necessary information to display on screen the original print edition. The second token in this line, namely 'a₃-da-um^{tug2}-II “(some kind of textile)”, is a sort of compound term or nominal syntagm, resulting from the juxtaposition of the lexeme 'a₃-da-um, the determinative for textiles^{tug2}, and the numeral II qualifying it (meaning a double textile?). Going back to the lacuna in front of it, as 'a₃-da-um^{tug2} textiles is always preceded by numerals, the broken sign may be further specified as ⑨, representing an unreadable digit. As for Line 3, the term spelled ib₂-III-dar-sa₆^{tug2} is exceedingly common in the Ebla texts. It can be analyzed as follow:

- ib₂ > lexeme for “belt”;
- III > a numerical attribute with adjectival function, possibly meaning “triple”;
- dar > an adjective meaning “coloured”;
- sa₆ > another adjective, meaning “of good quality”;
- tug₂ > determinative for textile items.

Within the Ebla corpus this term shows remarkable patterns of spelling variations, namely⁹:

1. ib₂-II^{tug2}-sa₆-dar (ARET 12 343, r.1,1)
2. ib₂-III-sa₆^{tug2}-dar (ARET 15 52, v.9,1)
3. 4 ib₂-IV-sa₆-dar 5 ib₂-III-dar^{tug2} (ARET 1 1, r.3,6, on the final^{tug2} see below)
4. ib₂-III^{tug2}-dar-sa₆ (ARET VII 133 r.3,1)

⁹ For the sake of clarity, we slightly normalized and adjusted readings appearing in the edited volumes, which often depends on the author's personal preference and expertise, see above §1.

Note that the adjectives *dar* and *sa₆*, as well as the determinative *tug₂*, may be freely placed after the numerical attribute attached to *ib₂* (either II, III, or IV). In addition, the determinative in example 3 above, although placed at the end of the line, apparently refers to both sets of belts. Obviously, this situation is a complicated one in terms of assyriological transliteration, as any alternative rendering remains problematic: for instance, a transliteration such as 4 *ib₂-IV-sa₆-dar* 5 *ib₂-III-dar-TUG₂* would imply that the sign *TUG₂* is not a determinative in the present context, but a sign having unclear reading within the sign sequence. As an alternative, a rendering 4 *ib₂-IV-sa₆-dar^{<tug2>}* 5 *ib₂-III-dar^{tug2}* would imply that the scribe simply forgot to write down the determinative associated with the first textile item, which is not the case.

These examples clearly reflect the fact that written language, being defined in the spatial domain, need not to follow the linear sequencing of spoken language, which is instead defined in the time domain. This fundamental property cannot be adequately represented in printed layout. This fact also impacts on text tokenization, as such process cannot simply rely on identifying strings delimited by boundary characters (whitespace, punctuation marks, or anything the user defined).

These issues prompted us to introduce two further steps in our digitization approach: 1) preprocessing of the individual texts prior to inclusion in the data repository; 2) handling of the smallest meaningful textual entities (sign-level, *infra*), regardless of their actual linear representation. These features, when paired to the annotation-level representation of the document (§3.4 and §4.3), may greatly help to enhance information, as well as to handle the problem of non-contiguous lexical entities.

Preprocessing is a fundamental step in the digitization process because it forces developers to formally address one of the central issues in the representation of all historical *corpora*: the tension between the need for a traditional print display – which most users are accustomed with – and the need for a heuristic annotation of all the relevant textual features. The primary input consists of text encoded according to a shallow markup system (BUCCELLATI 2011, with minor further implementations). A complete overview of rules and conventions adopted is not possible here. It suffices to say that special sequences of characters are used to mark entities at tablet, word, and sign levels (ex.: *a3-da-um-=TUG2-:2* represents ‘*a₃-da-um^{tug2}-II*’) (<http://ebda.cnr.it/> encoding). Broadly speaking, this shallow diplomatic encoding follows assyriological conventions, so that the enriched document still maintains a very high degree of readability (as opposed to standard XML annotated texts).

Special codes mark interesting features, such as: the physical condition of the tablet (e.g. *cb1* represents the indication “the beginning of column is broken”); broad semantic classes such as geographical or personal names (*g_* or *p_* respectively, placed in front of the actual term); the alloglottographic

system used in the actual spelling (e.g. °ma-°lik-°tum represents MA-LIK-TUM). The individual shallow annotated text files are then passed to a Python parser, which performs a series of preprocessing operations. First, it operates the necessary substitutions to represent the individual tokens according to their proper Unicode encoding. Second, it compares single words or group of terms against controlled dictionaries to attempt a preliminary linguistic disambiguation, as well as to check data consistency: the larger the collection, the larger the confidence. Third, it reads meta-linguistic notations (e.g. g_ for geographical names), consequently adding necessary attributes to the different occurrences of individual tokens, and cleaning word-level features off encoded properties. Finally, each token, alongside attributes and positional information, is passed to the data repository.

Turning back to the frequent and complex case of the “belts” described above, the parser splits the compound into different tokens and produces output such as the following:

- 1) ib_2 -III dar sa_6 -= tug_2 (ARET 3 42, r?.2.2)
- 2) 4 ib_2 -IV sa_6 dar 5 ib_2 -III dar -= tug_2 (ARET 1 1, r.3,6).

Such a solution is yet inadequate to represent the complexity of a writing system organized in columns and boxes. However, it is a good compromise. Each linguistic unit is represented by an individual token and this helps in maintaining the integrity of the original graphemic sequence.

But how is one to compare, for instance, patterns such as ib_2 -= tug_2 dar sa_6 and ib_2 dar sa_6 -= tug_2 ? The project’s innovative and original annotation system allows for the association of non-contiguous elements in a transparent way, as opposed to the opaque descriptive markup approaches. Because of its capabilities to reference multiple textual objects as a single entity and, above all, to work with overlapping textual objects, the EbDA system allows the user to annotate even arbitrary, non-contiguous portions of the document. As a consequence, also in the case of the problematic sequence 4 ib_2 -IV sa_6 dar 5 ib_2 -III dar -= tug_2 (where the final determinative actually refers to both items), it is possible to reference tug_2 as a proper token pointing to both objects, and then compare these derived “artificial” graphemic sequences to more frequent patterns in the repository where noun and determinative actually follow each other.

3.3 Sign-level

At the deepest hierarchical level, the parser also segments each text into its fundamentals building blocks, i.e. cuneiform signs. In this step, the parser checks the validity of input readings by comparing individual sign-level features against a list of known values (i.e. a sign list, a syllabary). At the same time, it performs a series of normalization steps. Each input value is thus converted

to the corresponding “sign name”. These are indexed by number and a series of attributes stored into the data repository. This level of representation of the documents allows for new text mining techniques in Assyriology, circumventing the issues of legacy transliterations (see above §2.6).

For instance, the graphemic sequence TUM 3 TUG₂ DAR ŠA₆ may be transliterated either $ib_2\text{-III}^{\text{tug}_2}$ dar sa₆ or $ib_2\text{-III}^{\text{tug}_2}$ gun₃ sa₆, depending on the author’s personal preferences, meaning “a triple colored belt of good quality”. The search engine is able to compare efficiently different sequences of characters (or strings) matching significant patterns based on their cuneiform “code points” and not the strings themselves. Thus, the search engine will consider $ib_2\text{-III}^{\text{tug}_2}$ dar sa₆ as equivalent to the $ib_2\text{-III}^{\text{tug}_2}$ gun₃ sa₆, despite of the fact that these two are actually rendered by different characters’ sequences. This search function is particularly powerful also in the case of sequences of signs of unclear interpretation. For instance, the sequence EN KA matches the personal names in Table 2 (they all reflect the same name; transliteration varies depending on author’s personal interpretation).

Results	Matching patterns by sign name
1) en-*KA-we-rum	>> EN KA PI AŠ
2) *EN-zu ₂ -*PI-*AŠ	>> EN KA PI AŠ
3) ru ₁₂ !!-zu ₂ -we-rum	>> EN KA PI AŠ
4) *EN-zu ₂ -*PI-*AŠ!	>> EN KA PI AŠ
5) ru ₁₂ -zu ₂ <-we?-rum	>> EN KA PI? AŠ

Table 2 – Output for the query of the sign sequence EN KA. The same personal name is transliterated in different ways (left column), depending on personal preference of the individual authors of the print volumes (ARET and MEE series).

3.4 Annotation-level

From a digital perspective, a document can be considered as the result of a set of nested logical structures, from the document itself to its smallest significant entities, which in turn may be arranged into parallel hierarchies of content objects (RENEAR *et al.* 1993). For instance, a digital cuneiform tablet may be represented at least by two overlapping hierarchies: the first one may be conceived as a physical representation of logical structures such as tablet > lines > signs or, as in the case of the administrative tablets from Ebla, a more complex structure such as tablet > columns > boxes > lines > signs; the second one may be conceived as a logical representation of a document, such as text > paragraphs > words. Parallel to these quite common hierarchies further levels

of information – and further hierarchies as well – emerge by addressing the peculiar nature of logo-syllabic writing systems (§2.4).

Actually, a document may eventually be annotated with any sort of meta-textual data arranged in categories relevant to the researcher (grammatical forms, translations, bibliographical notes, etc.), whereas more hierarchies may eventually emerge from practical contingencies of a given collection. This is obviously true not only for the Ebla texts, but also for cuneiform tablets from all periods of Mesopotamian history. For instance, in the case of legal texts dealing with real estates conveyances attested in several sites, one may be interested in defining further text structures, by dividing the document’s “operative section” – containing its nested legal clauses – from its contingency sections, such as the list of witnesses or the list of sold object(s). All this considered, a digital representation of a textual source should preserve all these information levels, be they determined by the original structure of the document, or derived by a set of post-processing annotations motivated by scholarly needs.

With this perspective in mind, in our digital collection a document is virtually not directly bound to any fixed hierarchy or structure. This is possible thanks to our original data model (§4.3), which allows to freely annotate any portion of the document, even overlapping and non-contiguous portions of text, thus generating logical structures also of arbitrary length and type.

4. A DIGITAL APPROACH

The complexity of the cuneiform writing system poses serious challenges to a digital representation of texts written with it. While the usual approach of adapting a markup language, like SGML and XML (GOLDFARB 1990; BRAY *et al.* 2006), maybe using a TEI-based encoding (TEI CONSORTIUM 2007), could be possible, at least in principle, we have chosen a different strategy. In fact, managing the different entities of the cuneiform writing system (like graphemes, signs, logograms, syllables, determinatives, etc.) would be impractical using such languages, which must already face several issues in representing complex texts, as discussed also in the recent literature (see for instance MAURIZIO, ORSINI 2010a; SCHMIDT 2012; BOSCHETTI, GROSSO 2014;). Among these issues, we can include the use of artificial solutions to represent multiple hierarchies within a document, such as stand-off markup or milestones; the growth of the document complexity when different levels of annotations are necessary; the difficulty of making queries that exploit the different linguistic levels of the documents and their annotations.

For all these reasons, our approach is a data centered one, based on the use of formal modelling tools and of database technology, through the following methodological steps:

1. First, we have defined a formal model of the cuneiform writing system and documents, with a notation based on the Unified Modeling Language (RUMBAUGH *et al.* 1999).
2. Then, starting from such a model, we have designed a relational database that includes the main elements of our model: this database is in fact a complete re-design of a previous database developed for the EbDA project, and in doing this we have adapted the model found in the previous step.
3. We have then populated the new database with data taken from the old one, that has been created during the last ten years by the researchers that have participated to the project. This work has been made in parallel with a significant clean-up of all data, performed through a set of sophisticated scripts. This work has produced a set of consistent data, stored in the newly designed database, which includes all the information recoverable from the previous one.
4. We are now developing a set of query patterns, which will be made available to scholars for exploring the data according to the new structure, by exploiting the different linguistic levels now available.

In the rest of this section, we will go through the details of such approach and the results that we have obtained so far.

4.1 *Notation used*

We use a graphical notation derived from that of UML to represent objects (RUMBAUGH *et al.* 1999; ALBANO *et al.* 2007). The notation aims to represent collections (or sets) of entities, called “classes”, with collections of association instances among them, called “associations”. All entities of the same class (belonging to the same set) have a common type, which describes which are the main facts of each entity in which we are interested in (the “properties” or “attributes” of the entity). Classes are graphically represented by boxes whose title is the class name and that contain the list of the attributes. One or more attributes, called Primary Key (PK), are declared to have different values for all the entities of the class.

An association instance describes a fact that relates entities from (possibly different) classes. For instance, since a tablet can be composed of one or more fragments, we define an association relating the class Fragments and the class Tablets. Associations have names, and possibly attributes themselves. They are graphically represented with an arc connecting the associated classes and decorated according to the different characteristics of the association.

The last conceptual tool that we use allows the representation of collections of entities seen at different levels of details (“specialization”). In many situations, entities that appear to have the same type from a certain point of view (therefore belonging to the same class), when analyzed more in detail,

show different sets of facts: for instance, certain subsets of entities have more properties than others, or are involved in associations peculiar to them. These subsets are called “subclasses”: a collection of entities is represented through a (super) class and one or more subclasses, meaning that entities in the subclasses are a subset of the entities of the superclass. The graphical representation connects subclasses to superclasses through hollow arcs.

4.2 A model for the cuneiform writing system of Ebla

With the use of this notation, we can now give a general model for the main aspects of the cuneiform writing system (Fig. 2). The model shows how a given sign – which is conceived here as a distinct graphic unit carrying information – relates to its possible material representation in context, i.e. an allograph (see above §2.5). In the schema, it is assumed that different allographs might appear in more than one sign repertoire currently available (sign lists and syllabaries, either including evidence from all periods of Mesopotamian history, or focused on specific text *corpora*). This implies a many-to-many relation between Allographs and Sign Repertoires. In addition, under the name of the relation (HasSigns) there are two properties, Code and Number, which belong to the relation itself, and not to either one of the two classes mentioned above (Allographs and Sign Repertoires).

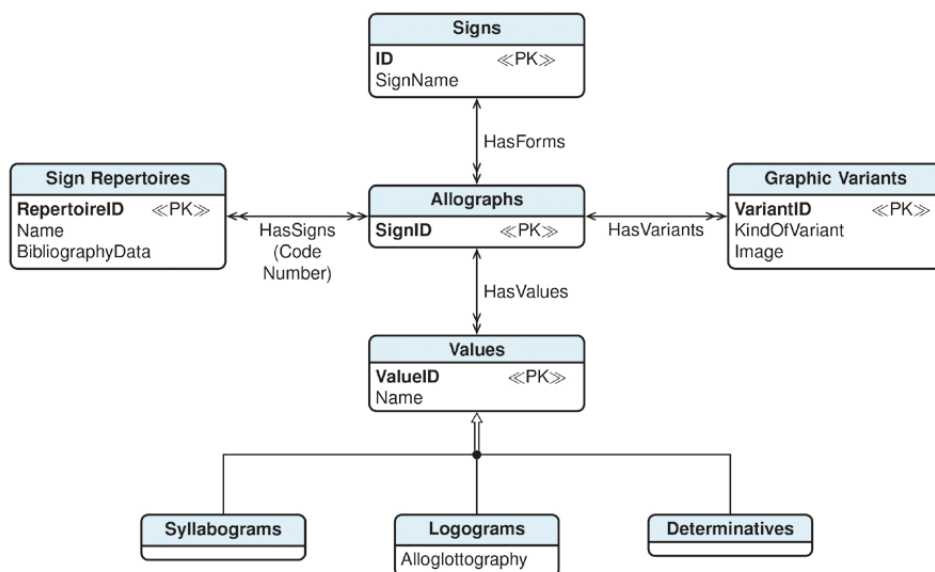


Fig. 2 – Model of the cuneiform writing system.

This is dictated by the fact that a given allograph may be listed in different sign repertoires according to different conventional numerical ids and/or sign name. Individual allographs usually appear in various palaeographical variants. The main property of the individual variants is the image or images of the variant itself, together with information related to it (for instance: geographical scope, scribal hand, period of attestation, etc.). The last important aspect is the fact that an allograph is a generic class of values that represents different alternative possible interpretations of it, through the subclasses syllabograms, logograms, and determinatives (see above §2.1-2.4).

4.3 The new EbDA Database

A re-engineering of this database and the related web application is now underway. In the rest of this section, we discuss the model of the new database through a conceptual schema using the notation introduced above. In addition, we shall show the relational database that has been developed starting from this model¹⁰. The aim of this new database is threefold: 1) it is meant to provide an updated digital edition of the entire *corpus* of transliterated texts belonging to the Ebla royal archives; 2) to capture the complexity of both content and writing system through *ad hoc* encoding system and a parser capable to populate with structured data the classes within the new data model; 3) to enhance the available textual information through a sophisticated annotation tool. Let us first show the model, and then discuss its different aspects (Fig. 3a).

Editions, Fragments and Tablets constitute the starting point of the database. Fragments (i.e. scattered pieces of cuneiform tablets) are the physical objects that have been recovered during different archaeological campaigns and consequently published in one or more scholarly editions. Note that a fragment can be the subject of several publications¹¹, while a publication can have as subject several fragments. A tablet is defined as a single fragment, or a group of fragments that have been recognized as belonging to the same tablet, that receives a unique identifier (TabletId). A tablet can be considered as an arrangement of areas, which could be either lines, or, in general, visually distinct parts of the tablet (columns, text boxes, borders, etc.). An area contains several occurrences of “tokens” (§3.2), or list of signs, which are alternative way of representing “words” (that is, the same word can be written in different epigraphic forms, and we are interested in representing the different occurrences of those forms in the areas).

¹⁰ Note that, for the sake of clarity, not all the details of the implemented database are shown.

¹¹ For instance, fragment TM 1975.G.01939 has been discussed in: NABU 1989/2: RitSucc, Text B (1992); VO 8/2, 3-11 (1992); Amurru 1, 36 and 125-128 (1996); and finally published as a fragment of ARET 11, 2.

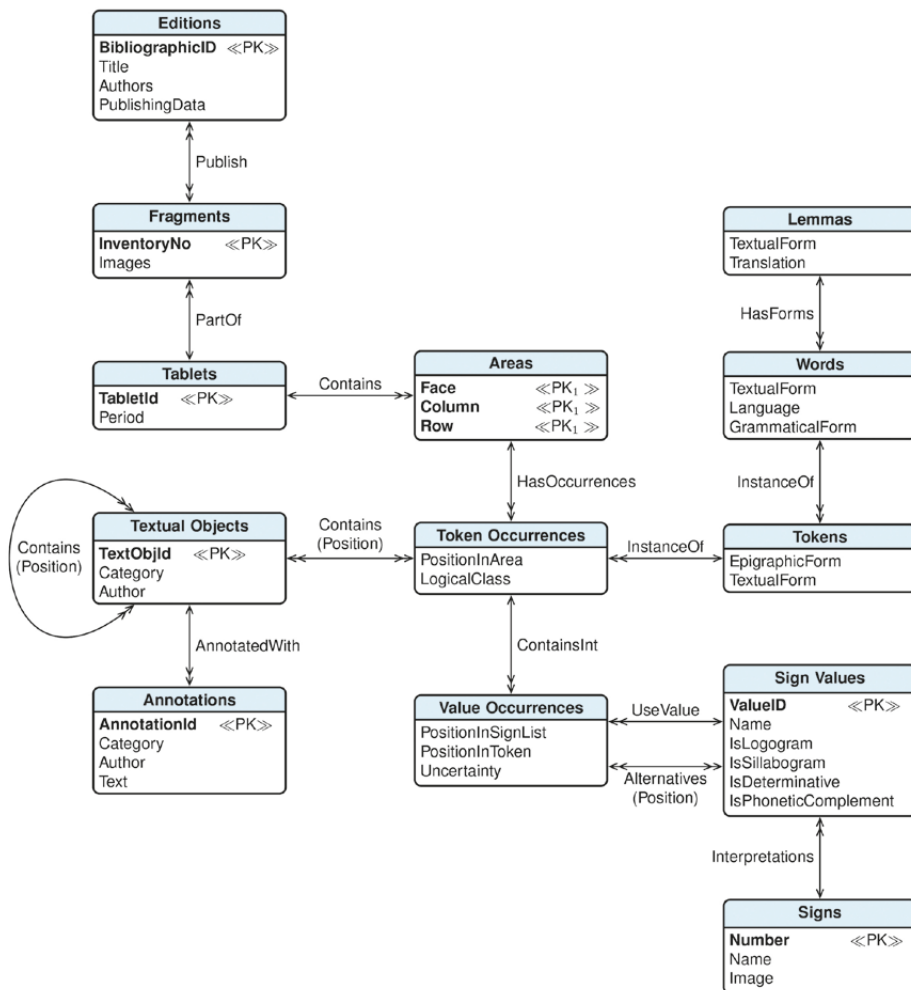


Fig. 3a – Conceptual schema of the database.

Each token occurrence has a position inside the area (the property PositionInArea) and a Logical Class containing its interpretation based on the actual context (for instance, as personal or geographical name). Note that in the Token Occurrences it is explicitly represented the essential information of the reading of each sign through the Value Occurrences class. An interpretation, related to a Token Occurrences, associates a main value to each sign in a specific position of the occurrence, together with possibly several alternative values, as in the case of the two allographs DIRI(SI.A) as opposed to DIRI(A.SI)

(both graphemic sequences SI.A and A.SI represent DIRI). Note the difference between the properties PositionInToken and PositionInSignList, which have been introduced in order to take into account the relationship of each sign value either within an individual token, or within the pertinent area (i.e. line of text). Note also the property Uncertainty, a number normalized in the range 0-1, which represents the level of confidence about the proposed sign interpretation. A Sign Value, to which an interpretation refers, has a name and a set of logical properties that describe if it is a logogram, a syllabogram, a determinative, a phonetic complement, and refers to the sign of which it is an interpretation.

Another section of the schema is devoted to the representation of the Tokens, that are the documents’ basic semantic components (§3.2). This is intended to record all unique forms in which a given sequence of characters may occur in a text transliteration, whereas the combination between the two classes Areas and Token Occurrences keeps track of all the instances repeated in the collection. The Tokens, thus, is intended to preserve all the unique occurrences of individual words together with all their epigraphic notation markers (e.g. parentheses, square brackets, etc.). At the same time, it collects unique occurrences of encoded strings not physically present on the actual text, but introduced by the editor(s) to preserve information such as the physical state of preservation of the source (e.g. the string [•-(•-•)], representing a broken textual case, having at least one sign missing at the beginning, possibly followed by two extra signs). This latter set of tokens are necessary to keep the digital representation of the document as close as possible to its printed layout, although they may hamper searching operations and comparison between terms.

This is the main reason for the introduction of the abstract object Words, which instead collects unique instances of individual words deprived of the epigraphical markers: thus, for instance, the two Tokens’ instances [i₃-]na-sum and [i₃¹-]na-sum (Tokens’s property EpigraphicForm) refer to the same Words’ instance i₃-na-sum (Words’s property TextualForm). The higher-level class of this branch of the scheme is represented by Lemmas, with its set of properties such as Translation. This latter class, Lemmas, is intended to archive headwords of the inflected words (i.e. canonical form or dictionary form), thus providing the project with the higher-level clustering property, able at exploiting the semantic richness of the collection. It allows indeed to perform search for headwords and, eventually, for their translations. For instance, the search for the verb *nadānum* or even for its translation “to give” will return as output both the logographic form i₃-na-sum (“he/she gives, they give/gave, given” etc.) and its syllabic variant *i-ti* that stands either for /yiddin/ “he gave” or /yiti(n)/ “give! (imp.)”. In other words, an individual object Lemmas will match two words (i₃-na-sum and *i-ti*), which in turn stand for:

	Words	Tokens	Token Occurrences
1)	<i>i₃-na-sum</i>	<i>i₃-na-sum</i> , [<i>i₃-</i>] <i>na-sum</i> , [<i>i₃</i>] ¹ <i>[na-sum]</i> , etc.	- <i>tokens indexes</i> -
	1 instance	23 instances (unique forms)	351 instances
2)	<i>i-ti</i>	<i>i-ti</i> , [<i>i</i>] ¹ <i>-[t]</i> , [<i>i</i>] ¹ <i>-t</i> , [<i>i</i>] ¹ <i>-t</i> [<i>i</i>]	- <i>tokens indexes</i> -
	1 instance	4 instances (unique forms)	13 instances

The last part of the schema concerns the possibility of storing different kinds of annotations over different parts of the text. This possibility is implemented by the classes Annotations and Textual Objects, patterned according to the Manuzio model to achieve as much flexibility as possible (MAURIZIO, ORSINI 2010a, 2010b). An annotation is characterized by its author, the kind of annotation (category) and the content, which presently is a generic text (in the future more complex annotations will be allowed). An annotation is related to a single Textual Object, which may represent any part of a text, including multiple, non-contiguous textual segments. A textual object is in fact a logical entity, which has a many-to-many relationship with Notation Occurrences (so that a textual object may include multiple tokens, and a token may participate to multiple textual objects). Moreover, a textual object can be composed of other textual objects as well, so that an annotation may be defined over complex structures. The properties of a textual object include its category (e.g. structural, grammatical, named entity, etc.) and the author of the creation of the object. Note that both the relationships Contains for Token Occurrences and for Textual Objects itself are characterized by a Position property, which allows a linear ordering over the constituents of an object.

4.4 The Relational Database Schema

The current implementation of the system is based on the PostgreSQL Relational Data Base Management System (RDBMS). The database has been built starting from the conceptual schema and transforming it in the relational data model used in the system. Basically, the relational data model represents data through relations, homogeneous sets of “tuples”, or “records” of elementary values. Often relations are assimilated to tables, but since relations are sets, they differ from normal tables in three crucial details: there are no duplicates, and no order is defined in the “rows” (the tuple) neither in the “columns” (fields or properties of the relation). Furthermore, a relation has a primary key, which is used to distinguish tuples, and to establish relationships among them, in the same or different relations (“referential integrity”: a tuple refers to another with an added attribute, called “foreign key”, that has the same value of the primary key of the referred tuple). For instance, in the following simple example we show a few tuples of the relations Tablets and Area:

Relation Tablets, the primary key is TabletID:

TabletID	Period
ARET 1, 1	reign of Išar-damu, Ibbi-zikir vizier
ARET 1, 2	reign of Išar-damu, Ibbi-zikir vizier

Relation Areas, the primary key is AreaID, while FkTabletID is a foreign key for Tablets:

AreaID	Face	Row	Column	FkTabletID
1	r	1	1	ARET 1, 1
2	r	1	2	ARET 1, 1
3	r	1	1	ARET 1, 2

We will show now the relational schema of the database by using a graphical notation to represent relations, their primary keys and their foreign keys. The notation is similar to the previous one, the main differences concern the lines connecting the tables: they have a different aspect since they show a direction (they start from the relation in which the foreign key is present), and their label which is the name of the foreign key. For instance, the example above about Tablets and Areas could be graphically represented as indicated in Fig. 3b. Note that, with respect to the conceptual schema, in Areas we have introduced both a primary key (AreaID), and FkTabletID, foreign key for Tablets, that connects any area to the relative tablet. For details on how to obtain a relational data model from the conceptual data model see, for instance, ALBANO *et al.* 2007; ELMASRI, NAVATHE 2015.

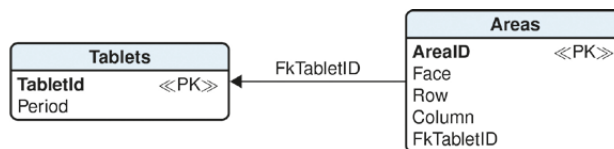


Fig. 3b

Fig. 4 illustrates the relational schema obtained from the conceptual one. The schema presented is a slight simplification of the current PostgreSQL database. This database, initially populated through a set of scripts that extract data from the previous versions of the system, is now the basis of the new web application, and it will provide the data to support a set of sophisticated data extraction and analysis operations behind those currently implemented. Here is a preliminary list of the planned operations:

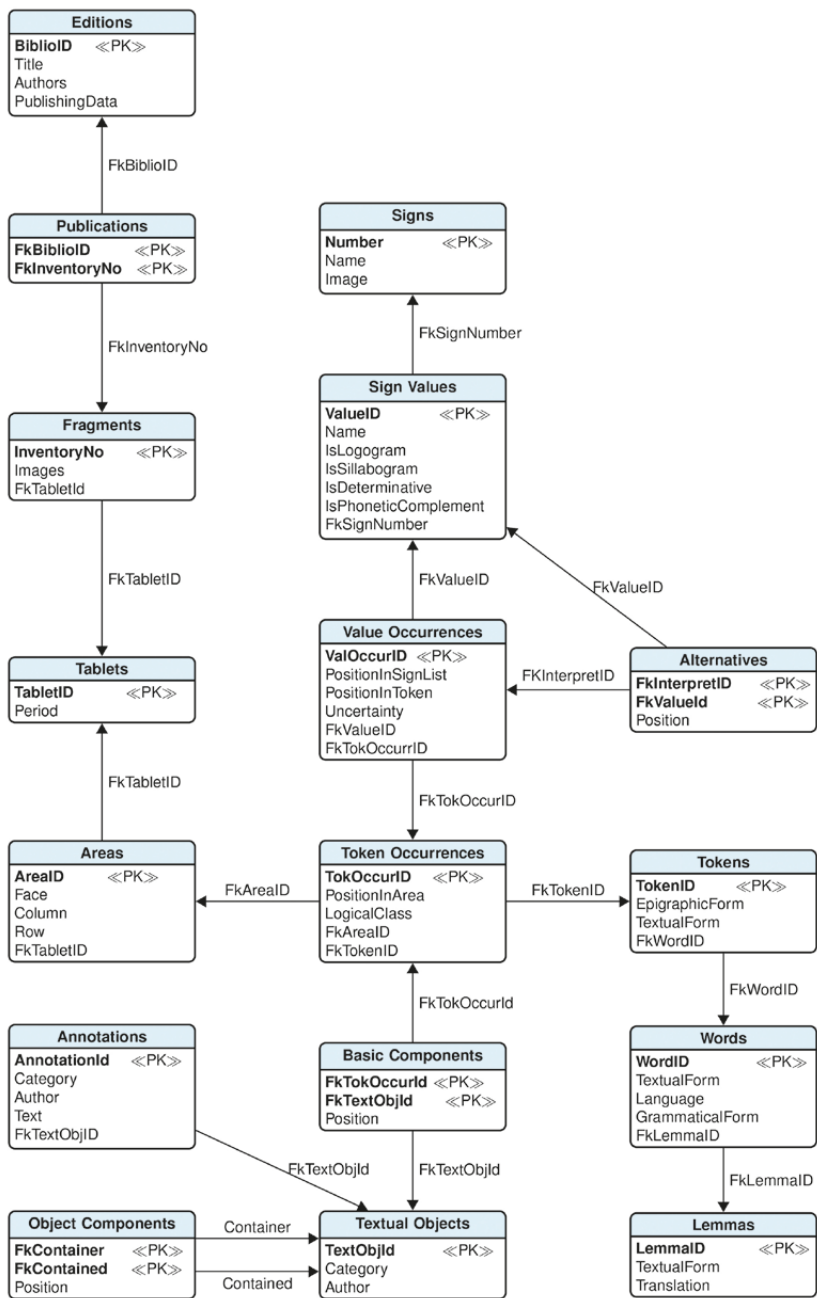


Fig. 4 – Schema of the relational database.

- Advanced queries based on regular expressions, matching any of the following: part of a word, whole word, word starting with, word ending with; user defined input string formatted according to PostgreSQL regular expressions syntax (<https://www.postgresql.org/docs/9.6/static/functions-matching.html>).
- Full Text queries (<https://www.postgresql.org/docs/9.6/static/textsearch-intro.html#TEXTSEARCH-DOCUMENT>) on English translations, based on stemming (e.g. a query for “goes” returns “to go” as well).
- Queries on lexical roots associated with the individual words attested in the Ebla documents, based on Lemmas.
- Queries for syntagmatic units: match one or more input strings within a user-defined word range: e.g. match the word for “house” (E₂) only when it is followed by the word for “king” (EN); match the word for “king” only when it is mentioned together with the word for “queen” within an interval of two words (e.g. “king and queen”).
- Co-occurrences: match texts containing an array of words (e.g. a list of city names, such as Ebla, Mari, Kakmium). This comes with a further option, namely an exclusion list (e.g. match all texts containing both Ebla and Kakmium, but not Mari).
- Queries for sign names: given an input reading, match all possible values attached to the corresponding sign. If two or more readings are passed as input, the query returns all words containing the corresponding input signs attached to them, regardless of their actual readings. Depending on user preference, the input string matches either two or more consecutive signs, or signs within a user-defined range.

5. FINAL REMARKS AND FUTURE PERSPECTIVES

The complexity of the cuneiform writing system of Ebla offers stimulating challenges to specialists in philology, information technology, and digital humanities alike. As Digital Humanities positively impacts on all fields involved in the study of the past, it becomes increasingly clear that traditional research methodologies must be matched by state-of-the-art research tools. The development of innovative software is however a slow and expensive process, as it requires close cooperation of experts in diverse fields. In order to minimize these drawbacks, it is important for philologists to develop hybrid expertise, which would greatly facilitate the dialogue with information technology experts. This would also greatly benefit their potential as scholars, as basic knowledge of query languages, scripting techniques, and databases opens up research avenues that would otherwise remain silent, not only because of the inevitable limitation in funding, but also because of the overall lack of vision. In other words, we are probably approaching a point where we should rethink the very notion of multidisciplinary. It is the fertile interplay of these newly

established scholarly domains that makes possible significant advancements in our understanding of the remote past.

We think that this paper is a first step in this direction, and we would like to stress which are its main contributions towards this end:

1. We have devised a first version of a formal model of the Cuneiform writing system: while this model will require certainly more refinements, it is a starting point for the design of a more sophisticated Knowledge base of cuneiform sources.
2. We have defined a database of the Ebla Archives that represents more information, and in a more useful way, of previous digital *corpora* of such texts. Through this database, state-of-the-art queries and analysis are possible, which may contribute to the advancement of our knowledge on this large and important set of documents, as well as on the fascinating world they witness.
3. In this database we have introduced an advanced annotation system, based on the concept of Textual Objects, which will support the work of the Ebla scholars through their collaborative work, enhancing the information contained in the database via annotations. In addition, we think that such mechanism is general enough so that it could be reused for other digital *corpora* of texts.
4. We have shown a few queries and analysis made possible by the new database. These should be considered as initial significant examples, as more analysis will be devised once the database will be complete.

FRANCESCO DI FILIPPO

CNR – Istituto di Studi sul Mediterraneo Antico
francesco.difilippo@isma.cnr.it

MASSIMO MAIOCCHI, LUCIO MILANO, RENZO ORSINI

Università Ca' Foscari, Venezia
massimo.maiocchi@unive.it, l_milano@unive.it, orsini@unive.it

REFERENCES

- ALBANO A., GHELLI G., ORSINI R. 2007, *Fondamenti di basi di dati*, Bologna, Zanichelli.
- ARCHI A. 1986, *The archives of Ebla*, in K.R. VEENHOF (ed.), *Cuneiform Archives and Libraries*, PIHANS 57, Leiden, Nederlands Historisch-Archaeologisch Instituut te Istanbul, 73-86.
- ARCHI A. 2003, *Archival Record-Keeping at Ebla*, in M. BROSIUS (ed.), *Ancient Archives and Archival Traditions: Concepts of Record-Keeping in the Ancient World*, Oxford, Oxford University Press, 17-36.
- BOSCHETTI F., GROSSO A.M.D. 2014, *TeiCoPhiLib: A library of components for the domain of collaborative philology*, «Journal of the Text Encoding Initiative», 8 (<http://journals.openedition.org/jtei/1285>; last accessed: 05/03/2018).
- BRAY T., PAOLI J., SPERBERG-MCQUEEN C.M., MALER M., YERGEAU F., COWAN J. 2006, *Extensible Markup Language (XML) 1.1*, World Wide Web Consortium Recommendation (<https://www.w3.org/TR/2006/REC-xml11-20060816/>; last accessed: 05/03/2018).
- BUCCELLATI G. 2011, *Digital edition and graphemic analysis of the Ebla texts*, in L. MILANO (ed.), *Archivi Reali di Ebla*, Edizione Digitale, 1, Cybernetica Mesopotamica, CD 4, Malibu, Udena Publications.

- CATAGNOTI A. 2012, *La grammatica della lingua di Ebla*, Firenze, Dipartimento di Scienze dell’Antichità.
- DEL FREO M., DI FILIPPO F. 2014, *LiBER: un progetto di digitalizzazione dei testi in scrittura Lineare B*, «Archeologia e Calcolatori», 25, 33-50.
- DI FILIPPO F. 2018, *Sinlequinnini: Designing an annotated text collection for logo-syllabic writing systems*, in A. DE SANTIS, I. ROSSI (eds.), *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, Berlin/Boston, De Gruyter, in print.
- ELMASRI R., NAVATHE S.B. 2015, *Fundamentals of database systems*, Boston, Pearson.
- GOLDFARB C. 1990, *The SGML Handbook*, Oxford, Oxford University Press.
- HUEHNERGARD J., WOODS C. 2004, *Akkadian and Eblaite*, in R.D. WOODARD (ed.), *The Cambridge Encyclopedia of the World’s Ancient Languages*, Cambridge, Cambridge University Press, 218-280.
- KREBERNIK M. 1982, *Zu Syllabar und Orthographie der lexikalischen Texte aus Ebla. Teil 1*, «Zeitschrift für Assyriologie und Vorderasiatische Archäologie», 72, 178-236.
- KREBERNIK M. 1996, *The Linguistic Classification of Eblaite, Methods, Problems, Results*, in J. COOPER, G. SCHWARTZ (eds.), *The Study of the Ancient Near East in the Twenty-First Century: The William Foxwell Albright Centennial Conference*, Winona Lake, Eisenbrauns, 233-250.
- MACKS A. 2002, *Parsing Akkadian Verbs with Prolog*, in M. ROSNER, S. WINTNER (eds.), *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Language*, Philadelphia, Association for Computational Linguistics, 3-8.
- MAURIZIO M., ORSINI R. 2010a, *A model and a language for large textual databases*, in S. BERGAMASCHI, S. LODI, R. MARTOGLIA, C. SARTORI (eds.), *SEBD 2010. Proceedings of the 8th Italian Symposium on Advanced Database Systems* (Rimini 2010), Bologna, Esculapio Editore, 254-265.
- MAURIZIO M., ORSINI R. 2010b, *Manuzio: a model for digital annotated text and its query/programming language*, in M. LALMAS, J. JOSE, A. RAUBER, F. SEBASTIANI, I. FROMMHOLZ (eds.), *Proceedings of the 14th European Conference on Research and advanced technology for digital libraries*, Berlin, Springer-Verlag, 478-481.
- RENEAR A.H., MYLONAS E., DURAND D. 1993, *Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies* (<https://www.ideals.illinois.edu/handle/2142/9407>; last accessed: 05/03/2018).
- RUBIO G. 2006, *Eblaite, Akkadian, and East Semitic*, in G. DEUTSCHER, N.J. KOUWENBERG (eds.), *The Akkadian Language in its Semitic Context: Studies in the Akkadian of the Third and Second Millennium BC* (PIHANS 106), Istanbul, Nederlands Historisch-Archaeologisch Instituut te Istanbul, 110-139.
- RUBIO G. 2007, *Writing in another tongue: Alloglottography in the Ancient Near East*, in S. SANDERS (ed.), *Margins of Writing. Origins of Cultures*, Oriental Institute Seminars 2, Chicago, Oriental Institute, 33-70.
- RUMBAUGH J., JACOBSON I., BOOCH G. 1999, *The Unified Modeling Language Reference Manual*, Reading (MA), Addison Wesley Longman Inc.
- SALLABERGER W. 2001, *Die Entwicklung der Keilschrift in Ebla*, in W. MEYER, M. NOWAK, A. PRUSS (eds.), *Beiträge zur Vorderasiatischen Archäologie Winfried Orthmann gewidmet*, Frankfurt am Main, Johann Wolfgang Goethe-Universität, Archäologisches Institut, 436-445.
- SALLABERGER W., PRUSS A. 2015, *Home and work in Early Bronze Age Mesopotamia: “ration lists” and “private houses” at Tell Beydar/Nabada*, in P. STEINKELLER, M. HUDSON (eds.), *Labor in the Ancient World* (International Scholars Conference on Ancient Near Eastern Economics 5), Dresden, Islet Verlag, 69-136.
- SCHMIDT D. 2012, *The role of markup in the digital humanities*, «Historical Social Research», 37(3), 125-146.

TEI CONSORTIUM (eds.), 2007, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, V. 5, TEI Consortium (<http://www.tei-c.org/Guidelines/P5/>; last accessed: 05/03/2018).
WELLISCH H.H. 1981, *Ebla: The world's oldest library*, «Journal of Library History», 16, 488-500.

ABSTRACT

The paper provides an overview of the digital tools developed as part of the Ebla Digital Archives Project, which aims to offer a digital edition of roughly 3,000 cuneiform tablets from ancient Ebla (modern Tell Mardikh, in western Syria), dated to the middle of the third millennium BCE. The Ebla archive is the oldest one in the history of mankind, for which extensive information concerning the primary setting of the documents is available. The archaicity of the writing system, combined with the inherent difficulties in reconstructing languages from the remote past (Sumerian, Akkadian, Eblaite), pushes us to rethink the strategies to properly digitally capture the complexity of these sources, of invaluable historical significance: administrative documents, literary texts, vocabularies, letters, etc. We tackled the problem through the development of a PostgreSQL database, which is populated by *ad hoc* Python scripts that parse input transliteration files, which in turn are encoded using a shallow mark-up language. The individual steps in such workflow are discussed, as well as the benefits in terms of advanced queries for information retrieval that such approach offers.

29
2018

€ 48,00

ISSN 1120-6861

e-ISSN 2385-1953

ISBN 978-88-7814-795-9

e-ISBN 978-88-7814-796-6

AC-29



ARCHEOLOGIA
E CALCOLATORI