Francesco Di Filippo

4 Sinleqiunnini: Designing an Annotated Text Collection for Logo-Syllabic Writing Systems

Abstract: Sophisticated writing systems, such as Cuneiform and Linear B, pose tremendous challenges for the development of digital corpora of annotated textual documents. The fact that both of them do not clearly represent the spoken form of the underlying languages, as well as the multi-level character of their logo-syllabic writing systems, has required the setting up of an *ad hoc* solution for complex data handling, aimed at capturing all of their features. While the usual approach of adapting a mark-up language would have been possible at least in principle, Sinleqiunnini relies on a different formal model, having been conceived from its beginning as a database driven framework. Such a solution was demonstrated to be more efficient than mark-up languages in representing parallel, overlapping hierarchies, while it also simplified prototyping of a set of complex queries to exploit the different information levels of these texts. Finally, it provided a more functional instrument to perform multi-user/multi-level annotation.

Keywords: Cuneiform, Linear B, relational model, XML schema, mark-up languages

4.1 The Project

Sinleqiunnini aims to be a software framework for the management of digital repositories of epigraphical sources, primarily concerned with logo-sillabic writing systems from the eastern Mediterranean basin, and their dissemination through the World Wide Web.[1]

The early stage of the project, which originated in 2006 under the supervision of C. Zaccagnini at the University of Naples "L'Orientale", focused on the setting up of a digital repository to store, visualize and query a textual database of cuneiform tablets from Emar (Syria), which was encoded as pure text files in the late '80s.[2] During these early developments, the project's primary objective was the setting up of a digital representation of texts to fit in the current transliteration methodology in order to

1 [www.pankus.com].

2 [http://virgo.unive.it/emaronline/cgi-bin/index.cgi].

Francesco Di Filippo, Consiglio Nazionale delle Ricerche, Roma

render the digital edition of texts virtually indistinguishable from their original printed layout. Since its beginning, it has been conceived to be Unicode aware and it was also one of the first projects dealing with cuneiform sources that bypassed the workaround of reproducing online editions through special, often unreadable, complex ASCII pseudo-encoding. This choice forced us towards high-level technical solutions that could efficiently manage variable-width length character encoding such as UTF-8. In 2006 the choices were quite restricted, so that we confidently relied on MySQL for data persistent storage and Perl as scripting language. At that time, the data model was a simple collection of occurrences of all lexical entities having been earlier tokenized from pure text files. Yet, even at this early stage of development, the software was quite efficient and responsive. Besides the capability of representing texts as in their printed layout, this first project already allowed search by string matching and regular expressions, in order to extract meaningful patterns by context, syntagmata and co-occurrences, and to produce glossaries of the digital collection.

Over time, Sinleqiunnini developed intermittently until it faced new, specific challenges arising from two very different textual collections. In 2011 the framework was employed to develop LiBER (Linear B Electronic Resources), a CNR project realized in collaboration with M. Del Freo, which aimed at producing a digital edition and a query tool for the Linear B documents (Del Freo & Di Filippo, 2014).[3] During this phase, Sinleqiunnini's architecture expanded with additional modules that were introduced in order to address issues concerning the spatial distribution of epigraphic phenomena, thus modifying the earlier data model architecture by enriching the system through Web–GIS capabilities.

In 2015, the data model architecture underwent a radical restyling. From 2008, the project was already in use for the management of the digital edition of the entire corpus of cuneiform texts belonging to the Ebla royal archives, EbDA (Ebla Digital Archives), a project of University of Venice "Ca' Foscari" in collaboration with L. Milano and M. Maiocchi.[4] During this last project development, we benefited from the extraordinary contribution of R. Orsini, who helped us develop a brand new relational scheme, the one in use today in Sinleqiunnini. This new implementation, which constitutes the object of the present article, has greatly enhanced database performances, while providing more effective querying and data-mining perspectives. More significantly, it also allowed the increase of the granularity of the database model, giving access to the management of the collection at its very basic unit level (i.e. cuneiform signs), and contributed to the design of a more consistent solution for multi-level/multi-user annotations (Di Filippo et al., 2018).

3 [http://liber.isma.cnr.it].

4 [http://ebda.cnr.it/].

4.2 Collection Design: Mark-Up Languages Versus Database Model

In a seminal article of 1990, with the purpose of designing a standard for encoding machine-readable documents, DeRose et al. (1990) boldly introduced the notion of "content object" as a logical structure of a document, "having to do with meaning and communicative intention". In the same contribution, they defined the document itself – i.e. its digital form – as a representation of an "ordered hierarchy of content objects" (the so-called OHCO model). In this view, a document is essentially the product of the juxtaposition of a series of nesting objects such as chapters, paragraphs, words, and so on, each of them containing elements of lower order. In the early '90s, this model was by far the simplest and most functional way to create, modify and format texts. Digital documents were represented in this way to support browsing, text mining procedures, and other sorts of special processing, and they were much more easily shared among different applications and platforms. It is not by chance, then, that this "ordered hierarchy of content objects" proved to be an effective premise in pushing the use of descriptive mark-up languages to represent digital documents. More specifically, it provided the most advantageous theoretical framework for projects involved in humanities computing, such as the Text Encoding Initiative (TEI).

Over time, however, some of the authors of the original thesis have identified a basic flaw in the apparent simplicity of the OHCO model. A textual document is indeed more often the result of several logical structures, a series of hierarchies that can also be reasonably considered "logical" (Renear, Mylonas, & Durand, 1993). By addressing the problem from different analytical perspectives, it soon emerged that a text may in fact have concurrent, overlapping hierarchies, and that this kind of textual source cannot be easily represented by a tree-shaped data structure. "Non-nesting information poses fundamental problems for any XML-based encoding scheme, and it must be stated at the outset that no current solution combines all the desirable attributes of formal simplicity, capacity to represent all occurring or imaginable kinds of structures, suitability for formal or mechanical validation. The representation of non-hierarchical information is thus necessarily a matter of trade-offs among various sets of advantages and disadvantages".[5]

Another important drawback in the adoption of a descriptive mark-up language for the architecture of a large repository of texts is deeply rooted in the metadata management. Any kind of information not directly belonging to any given hierarchy – i.e. extra-textual information such as metadata – can be tied only to the same structure of the text and must be expressed as a string of the mark-up language. This quite impractical restriction often pushes back-end developers towards the use of alternative data containers for persistent metadata storage. It is not uncommon to

5 [http://www.tei-c.org/release/doc/tei-p5-doc/it/html/NH.html].

meet mixed solutions indeed, solutions that pair mark-up languages for texts with relational databases for metadata. Such mixed workarounds are in use to such an extent that, as an apparent contradiction in terms, a giant of relational database management such as PostgreSQL since long (version 8.3) has been forced to introduce ways of storing loosely structured data like XML.[6]

Having discussed two of the main pitfalls in adopting a descriptive mark-up language in encoding textual sources, it is important to address more strictly some of the issues concerning the architecture of our project in relation to the peculiar type of sources it deals with.[7]

Consider, for instance, the case of a clay tablet, be it drafted through the archaic cuneiform of Ebla, or through the Linear B writing system. At least two concurrent, overlapping hierarchies may represent the structure of the document. There exists, in fact, a physical structure such as tablet > lines > words or, as in the case of the administrative documents from Ebla, a more complex structure such as tablet > columns > boxes > lines > words (see infra). These hierarchies overlap a further structure, that is the logical representation of the document such as text > paragraphs > words. This document has a title (e.g. MY Ue 652+656 or ARET 1 1), and may be enriched with information about the archaeological context of each of the fragments that constitute the document. This level of information (i.e. metadata) – although may be represented as a nesting structure as well (e.g. site > building > room > finsdspot) – does not belong to any of the hierarchies of the document and is far better represented by a relational model, whose ultimate goal is preserving data consistency and diminishing redundancies. This document may eventually be annotated, both with grammatical categories in the shape of tree-structured data and with commentaries made by different scholars, which over-time have given different interpretations and readings to some of the text's passages.

Parallel to this structure of our abstract sample text – a structure quite common of any digital collection of historical sources – further levels of information arise by addressing the peculiar nature of logo-syllabic writing systems. At the most general level, the documents of the digital collections considered here, record information by means of a rather large set of glyphs, usually ranging from a couple of hundred items up to a couple of thousand, depending on time, region, and text corpora. Within these writing systems, signs may be defined by functional classes that more or less

6 [https://www.postgresql.org/docs/8.3/static/datatype-xml.html].

7 During the past few years, we witnessed the emergence of a considerable number of projects involved in digital editions of cuneiform corpora (Charpin, 2014). However, notwithstanding relevant drawbacks in the use of XML based model even for modern alphabetic scripts, most of them relies precisely on the usual approach of adapting a mark-up language. As regards the Linear B writing system, instead, the problem concerning the management of annotated textual collections has been taken into consideration from a different perspective. Hitherto, the only two projects focusing on these sources rely on a database-driven approach (Del Freo & Di Filippo, 2014; Aurora, 2015).

correspond to the base classes of syllabograms, logograms and determinatives. In the case of Linear B, signs are usually more specialized than in the cuneiform writing system (Del Freo, 2016a). In the latter, a given glyph is often associated with more than one function, whose value sometimes can be revealed only by the context (Reiner, 1966). For instance, the "earth" sign, besides being used as a determinative for geographical names (usually rendered by superscript characters), may also stand for the word "earth" or "place" (respectively *erṣetu* and *ašru* in Akkadian), or for the syllabic value /ki/ of the personal name *a-bar-ki*. It is of note that the latter sign sequence deeply relies on its context: it can stand for either a geographical name (*a-bar*^{ki}, e.g. ARET 15 32) or a personal name (*a-bar-ki*, ARET 8 522, a name indicating a professional qualification). Furthermore, a given sign may be associated with more than one logographic value and also with more than one syllabic reading. For instance, the KA sign may be read ka "mouth", zu_2 "tooth", inim "word", etc., whereas the sign GA may be used to represent the syllables /ga/, /qa/, /ġa/, according to the so-called polyphony principle. Conversely, two or more different signs may end up having the same reading (homophony principle); in this case they are conventionally distinguished in modern transliterations by a lowercase numerical index (or accents). All this, at the more abstract level, entails the possibility of an unpredictable number of graphic variants of the same and unique word. From the perspective of the data model design, this also entails the necessity of enriching the structure of digital text with at least two further concurrent, overlapping hierarchies: the one bearing the actual reading of the text (i.e. its interpretation), the other recording un-interpreted graphemic sequences of conventional signs' names (Figure 4.1). In addition, scribal mistakes, signs added by modern editor, palimpsests, or erasures, often entail the addition of further nested structures for the management of the document's minimal units.

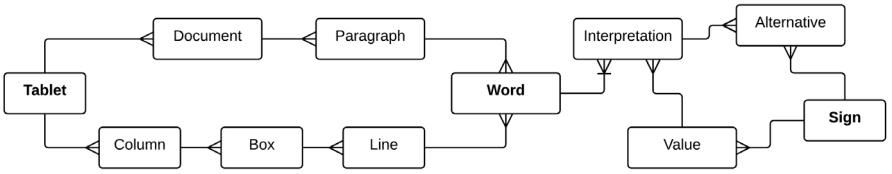

Figure 4.1: Synoptic scheme of parallel hierarchies

Another peculiarity of the cuneiform writing system is that graphemes, i.e. distinct minimal units within the sign corpus, may be arranged in a number of different ways, by: inclusion (partial or total), juxtaposition, ligature, crossing, and so on. For instance, the sign KU_2, which is used to express the verb "to eat", is composed by two graphemes: the sign for "food" (originally a pictographic representation of a vessel

for rations) either within the sign for mouth, or in close proximity to it. However, in the latter case an interpretation in terms of a different compound logogram, namely *inim gar* "to make a legal claim" (lit. "to place/put a word"), is also possible. This is an extreme case, but uncertainties in the interpretation of the documents may suggest leaving the readings of some sign or sign sequence open.

While the usual approach of adapting a mark-up language, like SGML and XML, perhaps using an EpiDoc-based encoding, could be possible at least in principle (DeRose, 2004; Smith, 2007; Witt & Metzing, 2010),[8] such a document would be very complex to create, manage and use. Even with sophisticated tools, an approach based on a descriptive mark-up language would require the development of *ad hoc* solutions, at the expense of greatly increasing the complexity of the representation and making even more difficult both the access to the textual collection and the processes of information retrieval (Iacob & Dekhtyar, 2005).

Some of the issues posed by logo-syllabic writing systems, to my knowledge, cannot be addressed by an architecture based on a descriptive mark-up language. The main limit in adopting, for instance, an XML scheme for a digital repository of transliterated primary sources from the ancient Near East and Aegean logo-syllabic scripts, in fact, concerns the impractical representation of complex, structured annotations. In XML, because attributes can only be represented by plain text and are directly bound to single elements of the hierarchy, it is virtually impossible to annotate non-contiguous portions of text. Moreover, annotations cannot overlap, nor is it possible to annotate concurrent hierarchies within the same and unique instance. These drawbacks in the annotation procedures of mark-up encoding systems, conversely, represent an essential prerequisite in the design of a digital collection planned for a variety of studies on the development of some of the earliest writing system in the history of mankind. Two examples will probably better clarify these very specific needs.

The layout of cuneiform and Linear B primary sources is very different from the literary documents for which mark-up systems were originally conceived. Linear B tablets do not pose many problems in this regard (Del Freo, 2016b). The writing system, moving from left to right by superimposed lines, more or less parallels a modern layout. The second millennium cuneiform system from Emar, at a great extent, arranges information in lines of even size and is also comparable to the modern stream of text, but some important exceptions may occur. For instance, on a consistent number of legal tablets of the Syro-Hittite scribal school (Seminara, 1998) the document ends with a series of seal impressions with Luwian hieroglyphic inscriptions, to which cuneiform legends bearing name and patronymic of the seal's owner are associated. The shape of these constructs is synoptically rendered as follows:

8 In the scope of cuneiform studies, it is worth to cite the ORACC project [http://oracc.museum.upenn.edu/].

.1	PN1 [seal impression]	PN3 [seal impression]
.2	son of PN2	son of PN4

As it is possible to observe, the natural stream of the text does not conform to the logical structure of the document one may want to annotate. Indeed, the primary level of information is not the relationship between the PN1 and PN3, which in fact lay one after the other, both on the original cuneiform tablet and in its digital reproduction. The main researcher's concern, potentially in order to analyse prosopographic ties, is actually the relationship between father and son, whose names are separated by the seal impression in the natural stream of the cuneiform tablet. As a result, in these cases, if one would like to save the original layout of the document, it would be necessary to annotate non-contiguous portions of text by putting together hundreds of these occurrences, a procedure that – even if possible in XML – would compromise human readability of the output and would require special tools to be fruitfully handled.

The second example of problems of practical annotation concerns a frequent feature occurring in the administrative archives of Ebla. The cuneiform writing evidenced in this site of inner Syria of the third millennium BCE shows many archaic features, especially in the layout of the administrative tablets, which form the bulk of the archive. Text is usually arranged in columns – to be read from top to bottom, from left to right – each containing several boxes, which in turn are inscribed with lines of uneven size. Each box usually contains a semantic unit such as, for instance, a number plus the noun it refers to, a verbal form, a preposition, and so on. In those documents, some items, which are very frequently evidenced (such as ib_2 "belt" that occurs more than ten thousand times), appear into a variety of "crystallized" graphemic sequences in which the order of the elements does not conform to any linguistic scheme. In this respect, the sequence ib_2-III-dar-sa_6^{tug2}, often transliterated with hyphens between each word unit (despite some variants that may occur, due to different editors' preferences), can be interpreted as follow:

- ib_2 > the main lexeme for "belt";
- III > a numeric attribute, probably denoting its length;
- dar > a qualifying adjective referring to the main lexeme, to be read "colored";
- sa_6 > a further adjective, meaning "of good quality";
- tug_2 > the determinative for this class of objects, that is "textiles".

There is more than one apparent incongruence in the way such a linguistic unit is represented. First, one would expect the three adjectives denoting the character of the "belt" (III, dar, sa_6) to be separated one from the others and from the qualified lexeme (ib_2) by means of a white space. Second, the determinative should immediately follow the lexeme and should not be placed at the end of the sequence: it determines the nature of the "belt" as a textile and not, as in this case, the last adjective in the chain

of morphograms. In short, this is a further example of the necessity of a flexible instrument to manage annotations of non-contiguous lexical units.

In addition, it should be stressed that exactly the same compound semantic unit is more often differently rendered, most frequently as a modern reader would expect (e.g. ib_2-II^{tug_2} sa_6 dar, ARET 1 1, r.4,7), but also by means of more convoluted sequences, such as: 4 ib_2-IV sa_6 dar 5 ib_2-III dar tug_2 (e.g. ARET 1 1, r.3,6). In this case, we have probably the clearest example of the difficulties in projecting these compound semantic units into linear patterns – which nevertheless is the common praxis for printed layout. In the latter example, the ancient scribe wrote the determinative for textiles (i.e. tug_2) at the end of the line (i.e. of the box), clearly with the intent of qualifying different semantic units of the same type with one determinative only. The typographic rendering of the sequence, the only viable option to preserve the integrity of the original document, however, has the counter-effect of generating a sort of linguistic ambiguity. Given the multi-level nature of the cuneiform writing system, an isolated sign tug_2 at the end of the line could even be considered as an independent linguistic unit. It would not qualify the nature of the preceding "belts" as textiles, but it would be considered as an independent word meaning "dress", thus distorting the sense of the whole sentence.

In order to better preserve the original textual layout and, at the same time, to safeguard the essential underlying level of information, it is then necessary to conceive a conceptual scheme that would allow annotations of non-contiguous lexical units and, as in the above-mentioned case, a system in which different instances can overlap without conflicting. In this regard, Sinleqiunnini allows annotations of any type, be they strings, structured values or references to other sections of the document. Being able to reference multiple textual objects as a single entity and, above all, to work with overlapping textual objects, it allows the user to annotate even arbitrary portions of the document. Going back to the above-mentioned sequence, the issue posed by this semantic pattern can be easily solved by referencing the determinative tug_2 twice as an instance of both the two preceding textual objects:

- original sequence:	4 ib_2-IV sa_6 dar 5 ib_2-III dar tug_2
- underlying annotation:	4 ib_2-IV^{tug_2} sa_6 dar 5 ib_2-III^{tug_2} dar

Finally, a further topic has been considered and technically solved by Sinleqiunnini's architecture. Logo-syllabic texts, as may be inferred by the writing system outlined above, are characterized by a substantial level of uncertainty and variation. Some text interpretations either rest on different scholars' readings, sometimes conflicting, or are still unavailable. Thus, the building of an integrated digital collection must envisage a multi-user, multi-level annotation system, in order to keep track of this set of overlapping interpretations. Yet, for the reasons discussed above, this system

cannot be easily handled by any of the mark-up languages commonly in use for the representation of large textual collections.

4.3 Sinleqiunnini Data Container

The project, since its latest development, greatly benefited by having been ported to Python as scripting language, Flask as framework web, and PostgreSQL as the relational data management system. In addition, in order to facilitate the conversion between incompatible type systems, the data container structure has been re-organized through an object-relational mapping system (ORM), namely SQLAlchemy (Myers & Copeland, 2015). This allows the project to interact with regular Python objects instead of working with database entities such as tables, documents, or Structured Query Language (SQL), yet it allows mixing use of the ORM with the SQL to satisfy very specific issues.

This substantial rewriting of the project source code has been the occasion to address, in a more formal manner, problems of data persistency by formalizing a new conceptual schema. This new schema has been deeply influenced by the high level of formalism of the Manuzio project (Maurizio & Orsini, 2010a), from which Sinleqiunnini differs in the structure of implementation (Maurizio & Orsini, 2010b).

Both the influence of Manuzio and the adoption of an ORM system have greatly contributed to the rethink of the nature of textual collections. Quite surprisingly, the above-mentioned OHCO model, which considers text as "ordered hierarchies of content objects", still proved to be the most serviceable theoretical framework for designing multi-layer textual documents. However, the many problems of adopting standard mark-up language solutions and, above all, the multi-level structure of non-alphabetic textual sources, led us to design an *ad hoc* solution for the management of these hierarchies of "content objects".

The core concept of our project's architecture is that a text can be represented as a set of hierarchies of either *textual* or *association objects*.

Textual object is an abstract representation of the different logic structures that contribute at defining a text as such; it has a logical meaning such as line, paragraph, word, sign, and so on. In other terms, a textual object is the sum of the portion of text with its structural (i.e. object's properties) and behavioural aspects inherited by ORM logic. Those aspects are of great help in maintaining data consistency. Textual object behaviour, in other terms, is a set of local procedures (i.e. methods), which help define computed properties, as well as perform operations on the represented portion of text. For instance, any time some text value is sent to the database, pertinent methods can check the validity of the information by validating it against a set of dictionaries (e.g. a syllabary) previously defined for the collection.

An *association object* has a slightly different nature, as well as a higher degree of abstraction: because textual objects cannot contain duplicates, association objects

are intended to keep track of positional and contextual information of textual objects. For instance, in our document collections a textual object "Tablet" is intended to represent an instance of a physical document, e.g MY Au 102, alongside all its attributes and references to lower-order logical structures. Obviously, being the highest order item in the hierarchy, there is only one instance of this type in each collection. Lower-order objects, anyway, need to be repeated as many times as the actual occurrences of these objects. The Mycenaean tablet MY Au 102 has references to 15 instances of the textual object "Line", to 35 instances of the textual object "Word", and so on. In other terms, a text may consist of many lines and a line may consist of many words: in relational model jargon, this is a typical example of a one to many relationship, which in turn is a perfect representation of a hierarchy by nested structures. However, it is also important to point out that in a textual document the same words may recur, sometimes with a very high frequency. In the case of our sample text, the Linear B logogram for "man" (by convention rendered by Latin word "VIR") appears 9 times in MY Au 102, thus representing more or less 26% of all the words of the above document. Do we really need to separately record each instance of the same word for "man"?

The relational model on which Sinleqiunnini rests allows a more convenient way to keep track of such information. The two textual objects, "Line" and "Word", were conceived to reference each other through an association object (i.e. "Occurrence"), which in turn is intended to permanently store the position of each of the unique occurrences of lines and words. In other terms, the logogram VIR exists only as a unique instance of the object "Word", but its position within the lines of the tablet is duly recorded as a numeric index by the Occurrence association object. The latter, moreover, is conceived to collect all the contextual attributes of its referenced textual object, attributes that may characterize the nature of the underlying finite sequence of characters (i.e. string) at a given position in the document.

The same is true for very frequent terms, as in the case of the Eblaite word *ʾa$_3$-da-um* (i.e. some kind of cape). To better illustrate this example, it is necessary to introduce a further characteristic of our data model. Sinleqiunnini's hierarchy of words rests on the difference among epigraphic notations ("Notation"), words ("Word"), and lexical entries ("Lemma"). The first textual object is intended to record all the possible, different forms in which a given term may occur, accounting for those signs not belonging to the original text and introduced by the critical edition to preserve a level of information concerning the physical state of the source (e.g. the square brackets for fractures). Of course, these characters are necessary to keep the digital representation of the document as close as possible to its printed layout, but they may hamper searching operations and comparison between terms. This is the main reason for the introduction of the abstract textual object Word. This collects unique instances of words from which all the editorial markers have been removed: thus, the two notations *ʾa$_3$-da-*[*um*] and *ʾ*[*a$_3$-d*]*a-u*[*m*] refer to the same word *ʾa$_3$-da-um*. A third textual object, Lemma, is intended to archive headwords of the inflected terms (i.e. canonical form or dictionary form), thus providing the system with a higher-level

clustering property. Turning back to the above example of association object, thus, Sinleqiunnini stores the very frequent Eblaite term *ʾa$_3$-da-um* as instances of three different nesting textual objects. At a higher level, there exists only one instance of the Lemma *ʾà-da-um*, eventually enriched with a set of attributes for its translations into modern languages. Then, there are multiple instances of the Word textual object such as *ʾa$_3$-da-um-I* or *ʾa$_3$-da-um-II* (by praxis, the base term and its specific numeric attribute are always treated as a compound element); finally, there are several Notation instances, as many as the single occurrences of its epigraphic notation variants. For instance, for the term's transliteration *ʾa$_3$-da-um^{tug2}*, that is a single lemma instance, more than one hundred different epigraphic words exist, which in turn refer to more than five hundred notations of the Eblaite word for "cape" in the collection of texts currently available.

Such a level of simplification has a significant impact on textual collection management. Any time the philological and epigraphical research provides a new reading for a given graphemic sequence – and this happens quite frequently in cuneiform studies – it is sufficient to update a unique instance of the object at the word level in order to make this change propagate by cascading effect on the textual collection as a whole. Moreover, it has relevant consequences in terms of searching and pattern matching procedures: Sinleqiunnini's search engine has to process only one item for each user's query. This resulting object, being characterized by the principle of inheritance of the object-oriented language, however, is intrinsically enriched by all its relationships with referenced objects, as well as by all pertinent positional and contextual information.

A last, concluding remark focuses on prototyping a multi-user and multi-level architecture to provide the system with cooperative annotations capabilities. In Sinleqiunnini, given that the structure of the textual object is capable of referring to arbitrary portions of the underlying text, annotations can be attributes of any type, be they strings, structured values or references to other textual objects. From this perspective, annotations are logical structures that can also encompass non-contiguous sets of lexical entities and, unlike mark-up language approaches, can overlap without the risk of conflicting. In addition, the relational database architecture provides researchers with the most efficient background for the management of multi-level sets of annotations.

The fact that text readings often rest on conflicting interpretations of different scholars poses remarkable challenges as concerns the number and dimension of annotations to be collected for each textual entity. Consider, for instance, the following excerpt from the Emarite tablet RAE 202:

> ll. 13-14: *u$_3$ a-nu-ma ṭup-pa* [*š*]*a* E$_2$ ^{d}IM *ma-ri* // ^{t}*tar-ṣi$_2$-pi$_2$ il-t*[*a-qu*] (Arnaud, 1986) (the tablet of the temple of god Ba'al, the sons of Turṣipu have taken).

This cuneiform tablet has been the object of several studies. Over time, very different readings have been proposed for these two lines, deeply conditioning historical research. From our perspective, these concurrent levels of information entail the necessity of a flexible annotation tool, not least because it is not yet possible to select a preferential interpretation for this text:

1) u_3 *a-nu-ma ṭup-pa ša* E_2 ^{md}IM-*ma-<lik ma>-ri* // ^{f}*tar-ṣi*$_2$*-pi*$_2$ *il-t*[*a-qi*$_3$] (Durand & Marti, 2003)[9]
2) u_3 *a-nu-ma ṭup-pa ša* E_2 $^{<m>d}$IM-ma-lik! // ^{f}*tar-ṣi*$_2$*-pi*$_2$ *il-t*[*a-qu*] (Cohen, 2009)
3) u_3 *a-nu-ma ṭup-pa ša* ⌜E_2⌝ ^{md}IM-ba-ri // ^{f}*tar-ṣi*$_2$*-pi*$_2$ *il-t*[*a-qi*$_3$] (Yamada, 2013)

1) the tablet concerning the house of **Ba'al-malik, the son of Turṣipu has taken**.	2) the tablet concerning the house of **Ba'al-malik, which Turṣipu took, …**	3) the tablet concerning the house of **Ba'al-baru, Turṣipu has taken**.

These four readings (that one of the tablet's first editor and the three new interpretations) are, except one, the result of the juxtaposition of the same number of tokens. In the interpretation no. 1, indeed, the assumed omission of two signs led the authors to split the personal name into two tokens, thus altering the paragraph length. As a consequence, when single word readings are different, it is impossible to simply collect these variants as attributes of the base instance of a word-level textual object. In Sinleqiunnini, instead, all those interpretations are intended as discrete logical units and these units are referenced to a common textual object type "Paragraph". As a consequence, alongside the basic reading of the document (eventually the one proposed by the original editor), there exist at least three parallel interpretations that potentially can be selected in the web-based user interface. At the same time, different instances of word-level objects, down to the collection of the minimal unit (i.e. cuneiform sign), are also referenced to the different interpretations, so that it is possible to perform searches even for these parallel discrete logical units. The resulting output then will specify the provenance of a given lexical entity and, eventually, if this word is part of an alternative reading proposal.

Finally, via bibliographic references, each new reading proposal is intrinsically tied to different scholar's authorities, which may also help end-users select pertinent interpretations for highly controversial text passages.

9 The <> markers stand for a modern insertion of cuneiform signs.

4.4 Conclusions

The complexity of the logo-syllabic writing systems 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 instruments is, however, a slow and expensive process. It requires close cooperation of experts in diverse fields, which in turn rests on the creation of a common, cross-domain language in order to facilitate this interplay. In order to minimize these drawbacks, it is important for philologists - and more generally, for researchers of the human past - to develop hybrid expertise, which would greatly help this dialogue with the information technology world. This would also greatly benefit their potential as scholars, as basic knowledge of data management and scripting languages may open up lines of research that would otherwise remain unexpressed. This is due not only to the greater paucity of financial resources, but predominantly because of a lack of vision of this complex system as a whole. It is the fertile interplay of these newly established scholarly domains that make significant advancements in the understanding of our history possible.

It is exactly with this spirit in mind that the Sinliqiunnini project has developed, although intermittingly, during these last ten years.

The project has defined a data model capable of representing the complexity of logo-syllabic writing systems by storing more information (and in a more useful way), compared to previous digital corpora of such a genre. Likewise, through this system, more sophisticated queries and analysis are possible due to the fact that data can be re-aggregated, on a case-by-case basis, through specific "views", which may reflect more strictly the specific needs of a given line of research. This, of course, relies on the fact that our approach rests on database technology, and the fact that our data model is not directly bound to any given textual hierarchy. Conversely, in our system each hierarchy, each ordered juxtaposition of logical structures, has the reasonable claim to be the fundamental digital representation of the document. There is no more need to "simply to pick a single hierarchy as the 'real' document hierarchy, and flatten all other hierarchies" (Renear, Mylonas, & Durand, 1993). The numbers of these equally important nested structures can grow over time without any significant impact on previously created querying tools or on the coherence of the collection as a whole. Since the structure of Sinleqiunnini is thought as a modular sequence of Textual Objects managed by a relational database, and not as a mere text file, each structure of the document is separated from the others and new Textual Objects can be added. New, logical structures can enrich the digital collection, also new structures that may not have been foreseen during the design phase of the digital repository.

This is the reason why the system (although in this regard it is still in its early stages of development) has been able to introduce an innovative annotation system, capable of bypassing intrinsic limits of the XML schemes, which will support the

collaborative work of scholars, enhancing the information contained in the database via annotations.

Finally, our data model supports a set of sophisticated data extraction and analysis operations:

- 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;
- Full Text queries on English translations – based on stemming (ex: a query for "goes" returns "to go" as well);
- queries on ancient lexical roots associated with the individual words, based on the Textual Object Lemmas.
- queries for syntagmatic units: match one or more input strings within a user-defined word range – e.g.: match the word for "house" (E_2) 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. 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.

During the past few years, we witnessed the emergence of a considerable number of projects involved in digital editions of cuneiform corpora (i.e. Charpin, 2014). Paradoxically, the tremendous amount of work has been perceived as something considerably different from traditional printed editions. Part of the issue is related to the actual evaluation system for the research products, which in EU countries at least is not yet capable of adequately evaluating the impact of state-of-the-art online digital tools, which are *per se* research products. Another part of the issue may be related to the fact that current online projects show a very high degree of variability. Most of them opted for proprietary conventions for the digital representation of their contents, either adapting an existing mark-up language or setting up an original one. Despite some relevant results, however, this process has hampered one of the prerequisites of the scientific research, which is the possibility of sharing information and data among scholars. We believe that it is time for modern philologists to consider the significant need for adoption of a shared digital grammar (encoding, data model, platform, query tools), specifically conceived for the management of the complexities of the logo-syllabic textual sources. We hope our project may serve as a starting point

in the definition of such grammar, to be further refined in order to assess the specific points of interest of the individual projects.

Bibliography

Arnaud, D. (1986). *Emar VI.3. Textes sumériens et accadiens*. Paris: ERC.

Aurora, F. (2015). DĀMOS (Database of Mycenaean at Oslo). Annotating a Fragmentarily Attested Language. *Procedia - Social and Behavioral Sciences*, *198*(1877), 21–31.

Charpin, D. (2014). Ressources assyriologiques sur Internet. *Bibiotheca Orientalis, 71*, 331–357.

Cohen, Y. (2009). *The Scribes and Scholars of Emar: Ancient Scribal Education in a Late Bronze Age City*. Winona Lake, Indiana: Eisenbrauns.

Del Freo, M. (2016a). La scrittura lineare B. In M. Del Freo & M. Perna (Eds.), *Manuale di epigrafia micenea. Vol. I* (pp. 123–166). Padova: Webster.

Del Freo, M. (2016b). Classificazione dei documenti e regole di trascrizione. In M. Del Freo & M. Perna (Eds.), *Manuale di epigrafia micenea. Vol. II* (pp. 247–256). Padova: Webster.

Del Freo, M. & Di Filippo, F. (2014). LiBER: un progetto di digitalizzazione dei testi in scrittura Lineare B. *Archeologia e Calcolatori*, *25*, 33–50.

DeRose, S.J. (2004). Markup Overlap: A Review and a Horse. In *Proceedings of Extreme Markup Languages, Montréal.* Retrieved from [http://conferences.idealliance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html], 2017/11/20.

DeRose, S.J., Durand, D.G., Mylonas, E., & Renear, A.H. (1990). What is text, really? *Journal of Computing in Higher Education*, *1*(2), 3–26.

Di Filippo, F., Maiocchi, M., Milano, L., & Orsini, R. (2018, in press). The "Ebla Digital Archives" Project: How to Deal with Methodological and Operational Issues in the Development of Cuneiform Texts Repositories. *Archeologia e Calcolatori,* 29, 117–142.

Durand, J.-M. & Marti, L. (2003). Chroniques du Moyen-Euphrate 2. Relecture de documents d'Ekalte, Émar et Tuttul. *Revue d'assyriologie et d'archéologie orientale*, *97*, 141–180.

Iacob, I. & Dekhtyar, A. (2005). Towards a Query Language for Multihierarchical XML: Revisiting XPath. In *Proceedings of the 8th International Workshop on the Web and Databases (WebDB 2005)* (pp. 49–54). Baltimore, Maryland: Citeseer.

Maurizio, M. & Orsini, R. (2010a). Manuzio: a model for digital annotated text and its query/programming language. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Proceeding ECDL'10 Proceedings of the 14th European conference on Research and advanced technology for digital libraries* (pp. 478–481). Berlin: Springer.

Maurizio, M. & Orsini, R. (2010b). A Model and a Language for Large Textual Databases. In S. Bergamaschi, S. Lodi, R. Martoglia, & C. Sartori (Eds.), *Proceedings of the Eighteenth Italian Symposium on Advanced Database Systems, SEBD 2010* (pp. 254–265). Bologna: Esculapio editore.

Myers, J. & Copeland, R. (2015). *Essential SQLAlchemy* (2nd ed.). Sebastopol, CA: O'Reilly Media.

Reiner, E. (1966). *A Linguistic Analysis of Akkadian*. London - The Hague - Paris: Mouton & Co.

Renear, A.H., Mylonas, E., & Durand, D. (1993). Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. Retrieved from [https://www.ideals.illinois.edu/handle/2142/9407], 2017/11/20.

Seminara, S. (1998). *L'accadico di Emar*. Roma: Università degli Studi di Roma "La Sapienza".

Smith, E.J.M. (2007). Using LPath Queries to Annotate Corpora: A Case Study of Elamite and Sumerian. In P. Zemánek, J. Gippert, H.-C. Luschützky, & P. Vavroušek (Eds.), *Chatreššar 2007. Electronic Corpora of Ancient Languages. Proceedings of the International Conference Prague,*

November 16-17, 2007 (pp. 121–134). Retrieved from [http://usj.ff.cuni.cz/system/files/Smith-Ch-2007.pdf], 2017/11/20.

Witt, A. & Metzing, D. (2010). *Linguistic Modeling of Information and Markup Languages: Contributions to Language Tecnology*. New York: Springer.

Yamada, M. (2013). The Chronology of the Emar Texts Reassessed. *Orient*, *48*, 125–156.