CHAPTER 10

Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence

Miller C. Prosser

The Ras Shamra Tablet Inventory (RSTI) is a research project co-directed by Miller C. Prosser and Dennis Pardee.[1] A primary goal of the project is to create reliable digital editions of the texts in the Ras Šamra-Ugarit corpus within a research-database environment.[2] More than just a data store of texts translated from their ancient languages, RSTI serves as a tool for addressing research questions. To this end, the project also seeks to integrate archaeological data from the excavations at Ras Šamra, including published archaeological plans, grid and square systems, and any other information freely available. Using the Online Cultural and Historical Research Environment (OCHRE), we are currently adding and curating data with the help of various workflow wizards.[3] To add new data to RSTI, we begin with a standard text transliteration saved as a Microsoft Word document or in another common document format. We load this document into OCHRE, which uses intelligent functions to atomize the linear transcription into individual signs or letters.[4] As part of this process, the application validates these signs and letters to make sure there are no typo-

1 Miller C. Prosser is a Research Database Specialist of the OCHRE Data Service of the Oriental Institute of the University of Chicago. Dennis Pardee is the Henry Crown Professor of Hebrew Studies at the Oriental Institute. See <https://ods.uchicago.edu/rsti/>.

2 Most people are familiar with the idea of a database. A research-database environment expands on the core structure of the database with tools and other features to help users work with their data.

3 A workflow wizard is an interactive database tool that guides the user through a series of common actions. For further information on OCHRE, see <https://ochre.uchicago.edu> (accessed May 1, 2017).

4 Atomization refers to the process of dividing data into many individual database items. A text is atomized into many database items, each of which represents a single grapheme, either a letter or a logosyllabic sign.

graphical errors.[5] Once the text is added to the database, analytical wizards guide the user through the tasks of finding words in project dictionaries, adding grammatical properties to the words, and identifying people and places in the texts. The importation and curation steps both employ processes developed specifically for the task of knowledge representation of philological data.

In this chapter, we explain the OCHRE data model and many of the tools developed for textual analysis. At points the discussion turns to the technical and may sound overly complex. However, it is important to remember that most of these complexities are hidden from the user. A typical user need not understand the logic behind a complex query. They need only know that they are able to search their entire corpus for texts that attest a specific phrase, and that the query will take into account all possible grammatical variants of the words in the phrase. This approach has evolved over the course of more than a decade to address some of the most complex writing systems from the ancient world.[6] In a sense, the underlying data in OCHRE is complex, but no more complex than the written language, no more complex than the original text, and only as complex as necessary to address the research problem. From the users' perspective, the texts look very familiar, and the complexities live mostly under the surface, hidden behind familiar-looking views of the text.

A Brief Introduction to Ras Šamra

The archaeological site of Ras Šamra is located on the eastern Mediterranean, near Lattakia, Syria. The ancient name for this site—and for the surrounding kingdom—was Ugarit. Ugarit was well situated, with access to trade routes, both land and sea, and to arable lands. The site was occupied almost continuously from the Neolithic period (c. 7000 BCE) through its eventual destruction in the twelfth century BCE during a regional period of instability. Later, Greek, Persian, and Roman garrisons occupied the site.[7] The historical significance of

5 OCHRE checks the text against a list of all known letters and signs in our ancient languages. If it finds a letter or sign that it does not recognize, it then warns the user that there may be an error.

6 The examples given in this chapter are taken from the languages and writing systems of the ancient Near East. However, the database and the various tools are appropriate for all languages and writing systems. Other projects using OCHRE are working with texts in modern languages such as English and German. Many of the complexities inherent in ancient languages such as Akkadian do not apply to modern languages, such as English. OCHRE is prepared to handle easier examples as much as it is ready to handle complicated examples.

7 Yon 2006, 15–18, 24.

the site was rediscovered nearly 100 years ago when a local herder happened upon underground tombs at the nearby harbor-side port of Minet al-Beida. The French Archaeological Mission was alerted and began investigations the following year. Excavations have continued from 1929 to the present, interrupted only by World War II and, more recently, by the Syrian civil war.[8]

In the centuries prior to its final conflagration, Ugarit was a cosmopolitan culture, exhibiting artistic and stylistic influences from Egypt, Mesopotamia, the Aegean, and Anatolia.[9] To date, Ugarit has yielded approximately 4,500 texts in various languages and writing systems.[10] The texts attest various genres, including texts conveying the economic concerns of the palace and ruling class, personal letters, accounts of ritual practice, and the famous mythological texts. This last genre has drawn a great deal of attention, as it provides an early example of a literary and religious tradition that is attested in a later form in the literature of the Hebrew Bible.

The scribes of Ugarit used a newly invented alphabetic writing system consisting of 30 letters. These same scribes were also trained in the Sumero-Akkadian logosyllabic writing system, which includes hundreds of characters in the local dialect.[11] Like the logosyllabic Mesopotamian writing system, Ugaritic letters are formed by impressing a stylus into a clay tablet, creating a series of wedges.[12] This type of writing is known as cuneiform, from the Latin *cuneus*, "wedge."[13]

8 Yon and Arnaud 2001, 7–8.

9 Matoïan 2008, 17–71; Akkermans and Schwartz 2003, 335–341.

10 Bordreuil and Pardee 2009, 8.

11 A logosyllabic writing system uses characters—or "signs"—that can represent words or syllables. For the Akkadian logosyllabic writing system, we call these word-signs logograms and the syllable-signs phonograms. Thus this type of writing is called logosyllabic, a combination of logograms and syllables. A third category of signs called determinatives plays a special role, sometimes marking grammatical information such as plurality, and sometimes marking the semantic category of a word. For example, the determinative ^{d} in the word ^{d}Aššur indicates that the word is a divine name. For the purposes of this discussion, it is only important to understand that scholars typically use font styles and formatting to differentiate the transcription of these three categories.

12 When a letter is written on a hard surface such as a stone object, the perceived outline of the wedge is scratched into the surface.

13 It is widely accepted that cuneiform writing was invented to express the Sumerian language (Woods, Emberling, and Teeter, 2010, 33). The writing system was later adopted by other cultures to express their own languages.

The Ras Shamra Tablet Inventory (RSTI)

At the Oriental Institute of the University of Chicago, Professor Pardee's work on Ugarit has now spanned five decades, over which time he has produced a substantial corpus of published and unpublished text editions. One of the first goals of RSTI was to create a framework within which to capture, preserve, and share this work. To be clear, RSTI presents a standard text edition with translation, epigraphic commentary, philological explanations, and interpretation. We present a recomposed view of each text that resembles a traditional print publication. However, we are also using various analytical wizards to add properties in order to create text editions that can be queried, summarized, and mined as data.

Among Pardee's published works, one of the first volumes transformed and ingested into RSTI was *La Trouvaille Épigraphique de l'Ougarit*, a volume written jointly with Pierre Bordreuil.[14] The goal of this volume is to provide a systematic accounting of all inscribed objects discovered at Ras Šamra and of Ugaritic texts discovered at other sites. The printed volume consists of a year-by-year, object-by-object presentation of archaeological inventory numbers, find spots, measurements, and other descriptive details. Once imported into RSTI, this data provided the primary spatial outline of the site of Ugarit, from the broadest excavation area down to each specific find spot.

Objects, texts, and images are the three main categories of data in RSTI. The database has entries for 5,700 objects, mostly tablets, but also vessels, seals, axe heads, and other items. To date we have added 950 text transliterations. The project currently includes over 32,000 tablet photos and drawings. Because one of our primary goals is to create a corpus of reliable text editions, we are taking care to add transliterations that meet project standards. This means that we are beginning with text editions created from first-person inspection of tablets in the National Museums of Damascus and Aleppo and in the Louvre.[15] It is our vision that RSTI will become a digital publication platform for all of these text editions. In the end, a text edition will include information about the tablet: its find spot, dimensions, and other observations about its physical characteristics. The edition will also include a text transliteration, a translation where

14 Bordreuil and Pardee 1989.

15 I was fortunate to gain access to study the Ras Šamra materials in the National Museums of Damascus, Aleppo, and in the Louvre thanks to generous research grants from the University of Chicago and to permission granted by the joint Syrian and French Mission at Ras Shamra (officially titled in French, "Mission archéologique syro-française de Ras Shamra – Ougarit"). Over the course of four separate visits, I was able to study and photograph hundreds of texts.

useful, specific epigraphic commentary, commentary on structure and interpretation, and bibliographic references.

An Overview of the OCHRE Ecosystem

The OCHRE database runs on a server professionally supported by the Digital Library Development Center at the Regenstein Library on the University of Chicago campus. All core data is stored in this database.[16] Users access data through the OCHRE Java application client. This client interface is a Java Web Start application that launches on any computer with an internet connection.[17] Because OCHRE is an online-database environment, project members from anywhere in the world have access to data live and in real time. If a user in Europe edits an item, users in North America immediately see these edits. The database, through the mediation of the OCHRE application, communicates with various external web servers and external databases. Project resources such as digital images, PDFs, and other supporting files are stored on these external servers and accessed for viewing and manipulation in the OCHRE client.

OCHRE data can be sent to external programs through an API.[18] For example, core OCHRE data can be sent to an R Server for statistical analysis and visualization, then returned through the database to the OCHRE client for viewing.[19]

16 RSTI is an OCHRE project. There are various other projects working on philological, archaeological, and other types of research in OCHRE. Projects vary greatly in size, from a few researchers to a network of international institutions. Each project has a discrete set of data that is unavailable to other projects by default. Project data is made accessible through credentials that are specific to users and projects. All data is stored securely and backed up on University of Chicago servers.

17 Java Web Start: a technology that allows one to launch a computer program without going through an installation process. For more information, see <https://ochre.uchicago.edu/page/java-user-interface> (accessed May 1, 2017). Java was chosen as the development language because it offers powerful features and allows deployment on various operating systems.

18 Application programming interface (API): a set of rules and tools that defines how computers can interact. In this context, the OCHRE API defines what database items are available to extract and send to other programs.

19 R: an-open source programming language used for statistical data analysis. See <https://www.r-project.org> (accessed May 1, 2017). An implementation of the R software environment can be installed on a server and accessed remotely by OCHRE, saving the user from having to install and maintain R on their local computer.

An Overview of the OCHRE Data Model

One of the simple principles underlying the OCHRE data model is that each discrete meaningful unit of observation is a separate database item.[20] Each database item is stored as an XML file.[21] Items may represent things that vary greatly in scale and type. An item could be an entire archaeological site or a single seed discovered in the course of excavation. An item can be anything from a place, to a person, to an image, to a text, to just about anything observable or conceptual. As mentioned above, the process of dividing data into these discrete units of observation is called atomization. In the end, these units function much like atoms, coming together to form larger and more complex entities.

Millions of individual XML files are related to one another through a variety of organizational methods, with the primary among these being the hierarchy. For archaeological contexts, the hierarchy expresses the relationship between broad spatial areas and more specific areas contained therein. To take an example from RSTI, the database item that represents the site of Ras Šamra stands in the hierarchy above, and contains, the area of the site called the Royal Palace.[22] The Royal Palace contains various rooms beneath it in the hierarchy. In these various rooms, we find many of the items that represent the inscribed tablets. In this way, the hierarchy organizes locations and objects into discrete areas into which all items can be contextualized spatially. As we shall see below, hierarchies play a central role in organizing textual data. In addition to organization through hierarchical relationships, database items can be linked in a wide range of ad-hoc and cross-cutting ways.[23] This approach, called the semistructured item-based approach, is in contrast to the class-based or relational data model in which similar items are stored in tables of columns and rows, then joined with other tables based on a common column called a key. In

20 All databases have an underlying data model, which is simply an abstraction that defines how data is connected and processed in the database. A data model is like a framework with rules. Applied to Archaeology, see in this volume, Matskevich and Sharon, 48.

21 XML stands for Extensible Markup Language and is a flexible data format. For further information, see in this volume, Bigot Juloux, 163–164.

22 For an accessible discussion of the site of Ugarit, see Yon (2006), specifically pages 36 and following for a discussion of the Royal Palace.

23 In the world of the digital humanities, the term "linked" can have a specific connotation, meaning the mode of modelling data for sharing across the web with other datasets with which one's data was previously not connected. Within the OCHRE system, the term refers to database items that "point" to each other, thereby creating a link between them. The database is a network of millions of files pointing at each other, i.e., linked to each other.

a semistructured item-based data model, each individual unit has a universally unique key and is free from the inherent restrictions of a table.[24]

The OCHRE Ontology and Data Types

We have just hinted at some of the ontological categories of data in the OCHRE data model. OCHRE employs what, in the world of information science, is called an upper ontology.[25] It is a highly generic, non-specifying schema of data categories applicable for use across many research domains. The categories of data types, such as Location, Person, Text, and Resource, are very broad. Each category of data is slightly different from every other, both in conceptual definition and in practical implementation. These data types are presented to the user as different hierarchies. What follows is a brief description of the data types that play a central role in RSTI.

The Locations & Objects data type is used for items that exist in space. Typically these are places and physical objects. These can be real places, observable in our current world or in excavated ruins, such as cities, neighbourhoods, streets, buildings, and rooms. These items can be movable objects that exist in space, such as an ancient coin, a modern book, or a clay vessel. Locations & Objects can be defined by spatial-coordinate data to indicate precise location in space.

The Persons & Organizations category is fairly self-explanatory. The primary unit of data in this category is either a person—living, deceased, real, or fictional—or an organization—anything from a publisher to any other conceptual group of persons. A person from this category of data is associated with the occurrence of their name in the text. In this way, a project can build relationships between names in texts. Properties can be added to each person to identify their familial connections, vocational roles, or any other piece of information that may help define them. In RSTI we are interested in identifying relationships of power among people of differing social strata. Below we explain some of the database tools that we have developed to aid in adding this type of information to the database.

---

24 Thuraisingham 2002, 155–171.

25 Schloen and Schloen 2014, <http://www.digitalhumanities.org/dhq/vol/8/4/000196/000196.html> (accessed May 1, 2017). For additional discussion, see in this volume, a) Bigot Juloux, in particular, 165–181; b) especially as applied to online publishing, Nurmikko-Fuller, 343, 348–350, 353–360; c) briefly, as applied to ontology in data-sharing in the archaeological field, Matskevich and Sharon, 47.

The Writing Systems category is foundational to all philology projects in the OCHRE database. Therefore, it is important to define some terms and explain the structure of this data type. A writing system is defined by a series of script units. A script unit is more than just a letter or a sign in the writing system. Each script unit is defined by various readings and allographs.[26] The English writing system would be fairly simple. There are only 26 script units, one for each letter of the alphabet. Languages from the ancient world are slightly more complex. In OCHRE we have created a writing system that represents the logosyllabic cuneiform writing system used in the ancient Near East. All research projects in OCHRE have access to OCHRE's standardized logosyllabic writing system. Like the list of letters that represents the English alphabet, this list represents all the known signs from ancient languages such as Sumerian and Akkadian. The list is standardized in the sense that it represents an attempt to create an architecture into which every sign from every logosyllabic writing system in the Sumero-Akkadian tradition can be recorded. In other words, the sign list is not divided into separate sign lists based on language or dialect.[27] On a practical level, the sign list is used for importing and performing automated processing of textual data at the data ingestion stage.

The Texts category will receive more attention in the following sections, but a few remarks will place this category into context alongside the other data types. As mentioned above, tablets and other objects are recorded in the Locations & Objects hierarchy. The object is not the same as the text inscribed on the object. The text is the series of signs used to communicate information. Inscribed objects from the Locations & Objects category are linked to the texts recorded on those objects.

The Dictionary is another category of data in OCHRE, and it has its own organizational structure. Dictionary data is highly integrated with textual data and plays another central role in RSTI. Any given dictionary is populated with lemma items. A lemma can be simple, with a basic definition and description, or it can be complex, with nested sub-entries of meanings. A lemma is further

26 A reading is a value represented by a sign. An allograph is a variant form of the letter, exemplified in most modern alphabetic systems by uppercase and lowercase letters.

27 The OCHRE sign list is based in part on Rykle Borger's *Mesopotamisches Zeichenlexikon* (2004) and supplemented as needed. In the next iteration of the sign list, we will add properties to the various signs and readings to indicate in which languages and dialects they are attested. This will allow the user to extract a list of signs that represent a dialect-specific sign list. We have published a version of this sign list online at <https://ods.uchicago.edu/signary/> (accessed May 1, 2017).

defined by phonemic forms, which in turn are defined by attested forms.[28] To explain a bit further, the Akkadian writing system allowed for variation in the spelling of a given word, even among words with the same grammatical properties. These various spellings are recorded in the database as attested forms. An attested form is a form of the lemma as it occurs in a text, represented by a sequence of signs. A phonemic form represents a grammatical interpretation of an attested form. In many cases, a given phonemic form is represented by many attested forms. To view this system as a hierarchy, an attested form is the lowest level of the hierarchical organization of items: lemma > phonemic form > attested form.

The Resources category organizes various supporting files, including images, PDFs, audio files, videos, GIS files, or webpages. Typically a resource will include an address where the file can be found, either on the web or on a project server. Most of these files can be viewed directly in OCHRE even though they are loaded from external servers.

Modelling Texts in a Database

To make philological data useful both to the researcher and to the interested public, it is essential to model the data in such a way that signs and words can be queried, referenced, and recombined easily and accurately. For English and some other modern languages, one can employ various computational methods that fall under the broad umbrella of Natural Language Processing (NLP) to analyze large blocks of text without first storing the text as tokenized units such as words.[29] Even if a text is not yet digitized, Optical Character Recognition (OCR) processing can usually produce a sufficiently accurate text. This process typically fails to identify a small percentage of letters in a printed document. However, this level of inaccuracy does not usually interfere with the analysis. Presently, and for the foreseeable future, OCR is not available for clay

28 Again, this complexity is less critical for modern languages; however, the highly variable and defective writing systems of the ancient world require these divisions.

29 NLP is a broad term that refers to methods for teaching computers to understand human language. At the risk of over-simplifying, the idea is that after a computer has processed a large number of words in a text, it should be able to extract meaning from a text it has never seen. Most work in this field has focused on teaching computers to understand English. The benefits of NLP have not yet reached the study of ancient languages. In this volume, see Svärd, Jauhiainen, Sahala, and Lindén (238–246), who focus on Pointwise Mutual Information (PMI), related to NLP, for Akkadian semantic research.

tablets.[30] In the meantime, we are left to determine the most effective method for handling ancient texts.

We propose a basic principle of best practice: textual data should be stored in the database in units that do not require further atomization before meaningful analysis can be performed. In other words, textual data should be stored as very small atomic units, either as words or as letters. How would one create a dictionary of attested forms when data is stored as a continuous text or as lines? One would be required either to duplicate words in another table that represents lemmata or to transform the existing data into discrete words. Both these extra steps are error-prone and cumbersome. In contrast, data that is highly atomized and granular, that is organized and described, can be accessed and recombined quite easily for multiple types of display and analysis. In essence, this is the power of a semistructured item-based data model.

As is the case with the other categories of data presented above, textual data in OCHRE is atomized into minimal meaningful units. When we think about texts from the ancient Near East, I think it is very clear that the minimal meaningful part is the grapheme (i.e., the syllabic sign or the alphabetic letter). There are many reasons to atomize texts into individual graphemes. First, many projects in our field are working on establishing reliable text editions, either as first editions of newly discovered texts or as re-editions of texts that deserve further attention. In this process, the project will require the ability to make comments and record metadata about every sign. Is the sign partially broken? Is it a scribal correction? Is it written above the line? From the perspective of someone outside the field, this may seem like an extreme degree of atomization. But from the perspective of a philologist, this is absolutely necessary. Therefore, each letter is its own database item. When the user wishes to view a text, OCHRE recomposes a view of the text based on hundreds or thousands of database items.

We have seen above how the hierarchy is a primary organizational structure for data in the OCHRE database. Textual data is also organized into hierarchies.

30 OCR: a process by which a machine scans a typed or handwritten document and converts the text into digital characters. For decades now, various attempts have been made to produce a system that can analyze digital images and identify cuneiform signs (Dirksen and von Bally 1997). No doubt advances will continue to be made toward this end. However, the variation evident across scripts and languages will make it difficult to achieve the level of accuracy currently available from OCR of printed English texts. If it becomes possible to capture even half of a cuneiform text accurately, this would save some effort on the part of the scholar. In the end, however, the scholar will still need to verify and correct the text edition because no level of inaccuracy is acceptable. For additional information, in this volume, see Eraslan, 296–297.

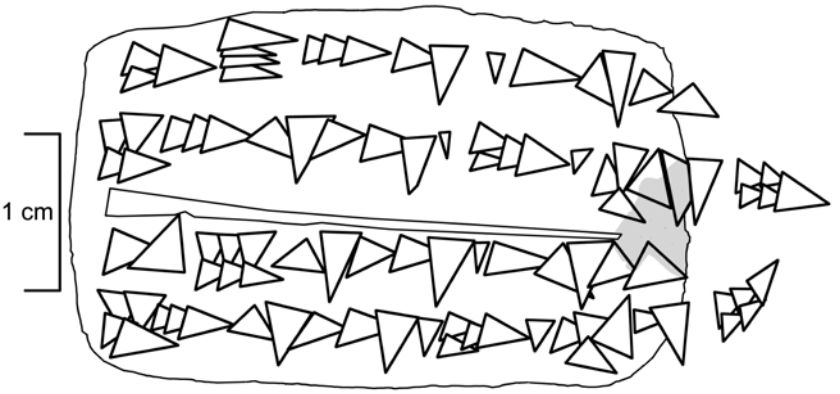

FIGURE 10.1 *Drawing of RS 3.320*

OCHRE maintains a conceptual distinction between the signs as they are visible on the object and the signs as they are interpreted as words by the scholar. The signs as they are visible on the object are organized into an epigraphic hierarchy, sign by sign. These same signs—as they are interpreted as words by the scholar—are organized into a discourse hierarchy. Any given word in the discourse hierarchy includes a list of links to the signs from the epigraphic hierarchy that compose that word. In this way, every word is associated with its constituent signs. This relationship between the grammatical form of the word and the attested spelling is leveraged to produce a dynamic dictionary.

Any typical text of 20 or 30 lines is composed of hundreds of database items. Taking a fairly simple example, the Ugaritic text RS 3.320 is a short text (Fig. 10.1).

It is composed of 36 signs and ten words. The text records two categories of cultic personnel (i.e., some type of priests or workers associated with a temple), each group with nine men and a donkey.[31] Following is the transliteration of the text I produced while studying the tablet in the National Museum of Aleppo.

31 The text is probably a census taken by the royal palace, but one may ask if the real-world events behind the text are best described as a conscription or a work-assignment. The information that we would need to reach this type of conclusion was omitted by the scribe because it was either obvious to all parties involved or, possibly, irrelevant to the situation.

Transliteration	Vocalization	Translation
(01) khnm . t⌜š⌝ʿ	(01) kāhinūma tišʿu	Priests, nine
(02) bnšm . w . ⌜ḥm⌝r	(02) bunušūma wa ḥimāru	men and a donkey;
---	---	
(03) qdšm . tšʿ	(03) q-d-šūma tišʿu	cultic personnel, nine
(04) bnšm . w . ḥmr	(04) bunušūma wa ḥimāru	men and a donkey.

The first word of the text is *khnm*, the masculine plural nominative form of the noun, which may be vocalized as "*kāhinūma*," "priests." [32] The four letters k-h-n-m are organized in the epigraphic hierarchy. My interpretation of the word as "*kāhinūma*" is organized in the discourse hierarchy, but it is also linked to the letters in the epigraphic hierarchy. To be clear, the letters that make up the word in the discourse hierarchy are the same database items found in the epigraphic hierarchy.

Recomposed Texts

Even though its textual data is atomized into individual letters, OCHRE produces recomposed views that look like text editions that are familiar to scholars. The image in Figure 10.2 is a screen capture from RSTI. On the left, we see the first part of the epigraphic hierarchy, expanded to show the individual letters. These letters are used to recompose the transliteration view of the text. The discourse hierarchy is used to create the recomposed view of the vocalized text (Fig. 10.2).

Textual data can be transformed very easily into other data formats, such as tables, PDFs, or word-processing documents. By creating a PDF directly from OCHRE, the user has an easy mechanism for publishing texts in a traditional paginated format. OCHRE can also publish textual data in a format appropriate for publication on the web. In this model, OCHRE data is published to an internal publications database in the OCHRE database environment, which in turn is made available to be accessed by a standard website. One final note on this

32 The Ugaritic alphabetic writing system indicates vowels only partially and only indirectly by use of three aleph signs: ȧ, ı̇, and u̇. Otherwise, no vowels were written (Bordreuil and Pardee 2009, §3.2-3.3) The vocalized form of the word is meant to convey the grammatical interpretation of the word.

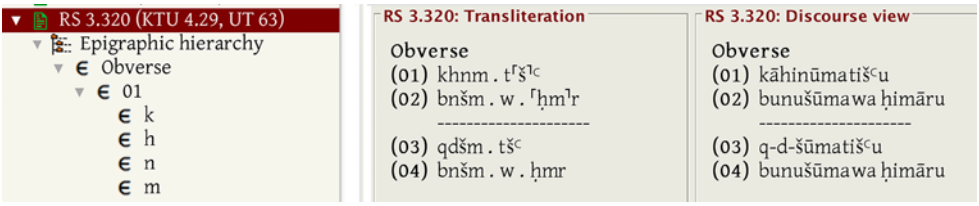

FIGURE 10.2 *Text, recomposed view in* RSTI

topic: OCHRE can also publish data in any digital standard, such as TEI.[33] Because OCHRE data is highly atomized, it is simply a matter of defining the desired format of the recomposed document.

Importing Texts

Textual data can be imported into OCHRE either from legacy formats or by typing directly into the database. Because OCHRE has been developed to handle all writing systems, the import process is highly flexible, customizable, and powerful.

OCHRE does not impose a transliteration system upon projects. Each project is free to customize OCHRE to understand the significance of lowercase, uppercase, italic, and superscript transliteration. For example, RSTI uses the following transliteration style for texts in which a logosyllabic script is used to record the Akkadian language:

> Lowercase italic = phonogram, Akkadian language
> Uppercase regular = logogram, Akkadian language
> Lowercase regular, superscript = determinative, Akkadian language
>
> For example, the following lines are a transcription of RS 17.238:11-12:
> (11) *šum-ma* DUMU^{meš} KUR *ú-ga-ri-it*
> (12) *ša* KUR-*ti ša-ni-ti*

33 The "Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics" (<http://www.tei-c.org> [accessed May 1, 2017]). For additional information, see in this volume, Bigot Juloux, 164–165.

During the import stage, OCHRE examines a transliterated sign, uses the specification to identify the type of sign based on formatting, then finds this specific reading in the appropriate writing system. Once imported, any given sign in a text is linked to a specific reading of a sign in a writing system. So, in the excerpt above from RS 17.238, the import process examines the first transliterated sign, "*šum*," and looks it up in the Sumero-Akkadian writing system. Because the sign is transliterated in lowercase italic, and because I instructed OCHRE that this format is used for phonograms, the query considers only phonograms in the writing system. It finds the value "*šum*" as a phonogram in the writing system under the sign "TAG." The sign "DUMU" is recognized as a logogram (a word sign), which in this case stands for an Akkadian word that means "son." The "*meš*" sign after "DUMU" is recognized as a determinative, which in this case marks the noun as plural, "sons."

For the sake of illustration, we will follow part of the text-import process for one RSTI text. The following text (RS 15.076) is a bilingual text. The recto, written in alphabetic Ugaritic, lists proper names, each with a number noun. For example, line one says, "*sākinu*, thirty." The verso, written in logosyllabic Akkadian, lists quantities of garments.

Recto
(01) *skn . tltm*
(02) *ỉytlm . tltm*
(03) *ḫyml . tltm*
(04) *ġlkz . tltm*
(05) *mlknʿm . ʿšrm*
(06) *mrʿm . ʿšrm*
(07) *ʿmlbỉ . ʿšrm*
(08) *ʿmtḏl . ʿšrm*
(09) *yʿḏrd . ʿšrm*
(10) *gmrd . ʿšrm*
(11) *ṣdqšlm . ʿšrm*
(12) *yknỉl . ḫmš*
(13) *ỉlmlk . ḫmš*
(14) *prṯ . ʿšr*
(15) *ỉbn . ʿšr*

Verso
(16) 3DIŠ TÚG^{meš} GAL
(17) 1ÀŠ TÚG^{meš} TUR^{meš}
(18) 2DIŠ TÚG^{meš} MÍ^{meš}

(19) 1IÁ TÚG GÚ.È^{meš}
(20) 1U TÚG^{meš} *ku-ub-šu*

There is no special encoding necessary to communicate with the import process. One simply supplies the import wizard with a block of text such as the one above, which looks like a standard text transliteration. In fact, this block of text was simply copied from a Microsoft Word document and pasted into RSTI. OCHRE has been trained and instructed on how to understand this document. The import process has been told that lowercase non-italicized text is alphabetic Ugaritic, that lowercase italicized text represents Akkadian phonograms, that uppercase text represents logograms, and that superscripted text represents determinatives. So, it finds "s" at the beginning of line one and understands it as a letter in the Ugaritic writing system. The import performs a query and finds "s" in the alphabetic Ugaritic writing system as a valid letter. By virtue of this link to the writing system, where we specify various properties, including Unicode encoding, OCHRE now knows that this letter can be represented as the Ugaritic cuneiform letter identified by Unicode point 𐎎.[34] At this point, the handling for the first epigraphic unit is complete. The import created an epigraphic unit called "s," linked it to the script unit "s" in the Ugaritic writing system, and placed the letter in the hierarchy that represents the Recto of the text in line 1. This logical loop continues through the end of the text, handling each section, line, word, and sign.

In addition to identifying the structure of the text and the value of each sign, the import process also creates words from the signs. It understands that a space indicates a word boundary. As such, the import creates a word, "*skn*," from the first three epigraphic units. Again, the individual epigraphic units exist in an epigraphic branch of the text's hierarchical structure. These same epigraphic units are linked to a single word in a separate discourse hierarchy in the text's hierarchical structure. In the end, every text is a network of signs and words.

34 We need our computers to communicate effectively with each other (and with us). On a very basic level, this means that computers must use an agreed-upon system of representing characters. The Unicode Consortium was created to promote this standard. One of the most significant contributions of this group is the creation of a standard that defines which underlying computer code is used to represent (nearly) every character in every writing system, even Ugaritic. The number of each character can be referred to as a Unicode point. For further information on Unicode, see in this volume, Eraslan 300–301.

Text Analysis Workflow Wizards

With RSTI, we aim to use our text editions to address various research questions. For example, we hope to identify socio-economic networks among the persons named in the texts. It is our working hypothesis that relations among persons of varying levels of power may be described using the terminology of patron-client relations.[35] To test this hypothesis, we are using OCHRE to help us identify the social positions of individuals named in texts. While this could prove to be a time-intensive activity, we have attempted to reduce the potential time needed by developing some database tools to help move the process along more smoothly. These tools, which we have been calling wizards, aid us in a variety of tasks that help us address this research question. There are currently three wizards for philological analysis: (1) the lexicography wizard; (2) the prosopography analysis wizard; and (3) the gazetteer wizard. Following is a brief overview of each of these workflow wizards.

Lexicography is the practice of compiling glossaries and dictionaries, a task we wish to perform in RSTI. Therefore, we developed a lexicography workflow wizard to link words in the text to lemmata in the dictionary, add parsing properties, and even add new words to the dictionary. To use the lexicography wizard, the user views a text in OCHRE, then launches the wizard (Fig. 10.3).

The wizard checks the words in the discourse hierarchy and finds the first word that is not yet linked to the dictionary. It searches the dictionary for attested forms that match the attested form of the word in the text. The wizard returns a list of all matches, allowing the user to select the correct word. If no match is found, the user may choose to add a new form to the dictionary while still in the workflow wizard. Once the new form is added to the dictionary, it is also linked to the current word in the text. At this point in the wizard, the user has the option to add grammatical parsing properties to the phonemic form of the lemma in the dictionary. These properties can include the typical range of nominal declension and verbal parsing properties such as person, number,

35 Patronage is one of the connective tissues of social integration joining individuals within and across the core institutional structures, such as kinship groups, vocational groups, and political authority. The ties between patron and client frequently augment pre-existing modes of social or economic protection such as kinship, vocational, or village ties. The relationship between a patron and his client(s) is based on a restricted access to the available productive resources (Eisenstadt and Roniger 1984, 171). This restriction does not represent an attempt to exclude a significant sector of the population from productive resources. On the contrary, the principle of restrictive access allows a wide range of individuals to gain conditional access to the available resources, primarily through the intermediary mode of exchange called the patron-client relationship.

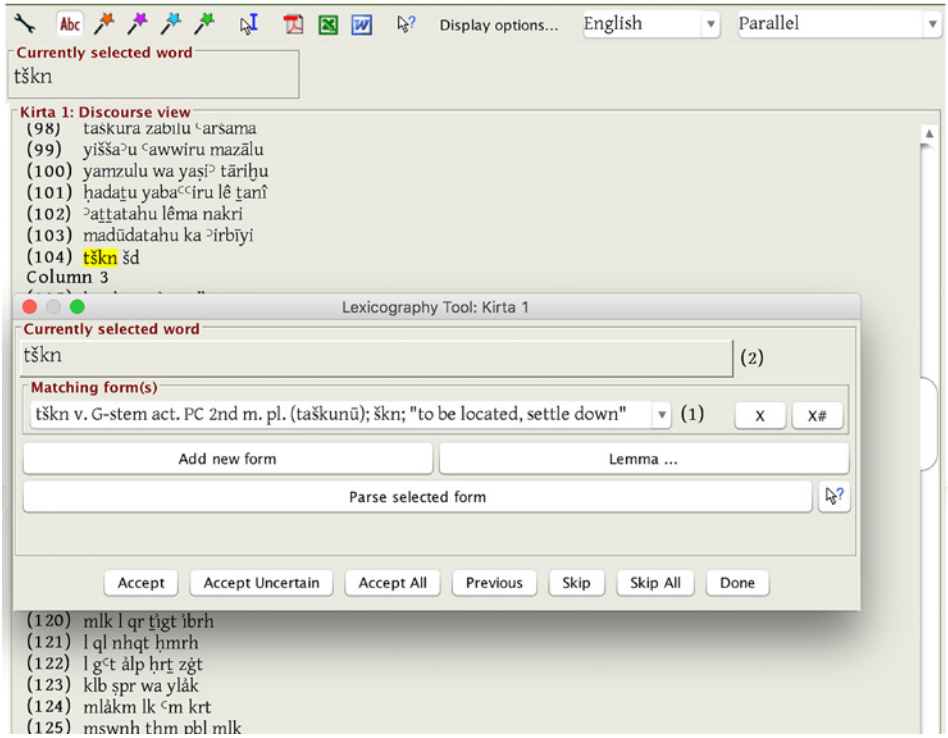

FIGURE 10.3 *The Lexicography Wizard*

gender, and case. Over time, the import process and wizard are likely to find more matches automatically. Essentially, the user is training the database to become smarter.

The lexicographic work on the texts yields a dictionary populated with live links to the attested forms in the texts (Fig. 10.4).

To review this network of data briefly, a series of epigraphic units forms a discourse unit (i.e., a word). A word is linked as an attested form to the dictionary. Attested forms are organized under grammatical forms. The grammatical forms in the dictionary have properties that indicate the parsing of the word. The grammatical forms are organized under the lemma heading. At the level of the lemma, a project can define various meanings, assign properties, and indicate cross-references or other details. This graph network of data is one of the primary characteristics of an item-based semistructured data model. New connections are made on the fly, not based on predefined table structures.

The second workflow wizard aids the user in adding prosopography properties to personal names in the text, specifically the types of properties that

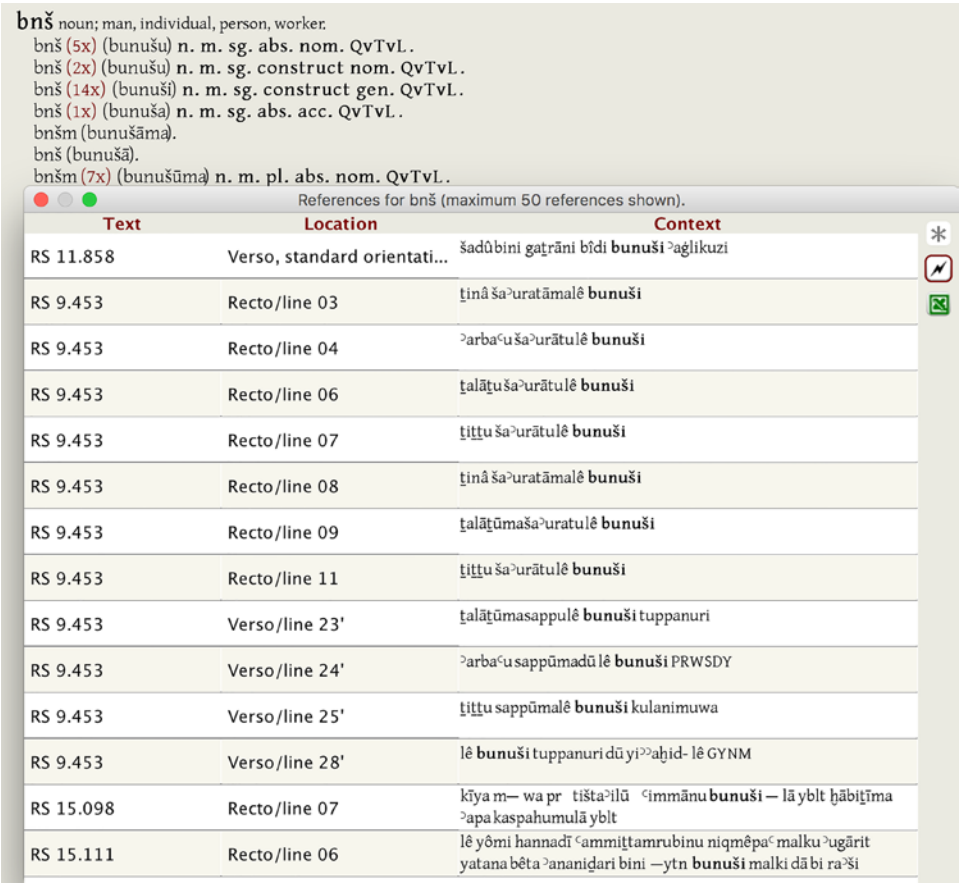

FIGURE 10.4 *Dictionary lemma in RSTI*

would help identify a person's social, vocational, and familial connections (Fig. 10.5).

Like the lexicography wizard, the prosopography wizard iterates over the words in the text and stops at the first word it recognizes as a personal name. The wizard returns a list of matching persons, allowing the user to select the correct person. This wizard uses the properties added during the lexicography process to determine if a word is a personal name. When the wizard finds a personal name, it performs a query of all known names in the Persons & Organizations category. Because names can be spelled many ways, and in different writing systems—many names are attested in alphabetic Ugaritic and logosyllabic Akkadian—the query takes into consideration all these spellings when matching against names in the Persons & Organizations category.

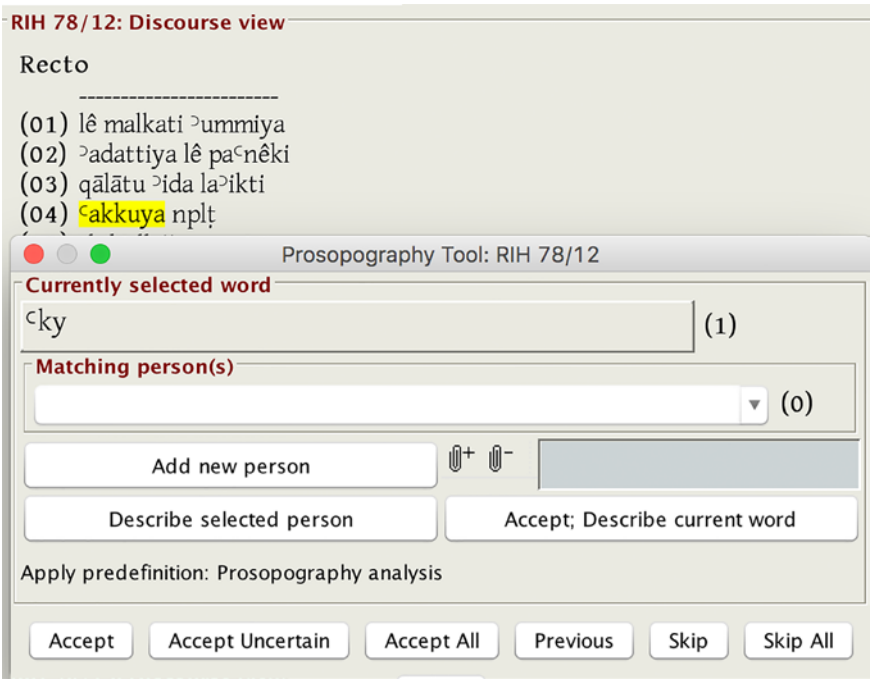

FIGURE 10.5 *The Prosopography Wizard*

Once enough names are identified and associated with persons, and enough relationships are defined, this data serves as the basis for social network analysis. Who is related to whom in real or fictive kinship relationships?[36] Who provides goods, services, or information? And what sort of hierarchical or other relationships of power can be detected in these connections?

The Ras Šamra texts frequently record the names of towns, villages, and estates that are located inside the kingdom of Ugarit or even farther afield. Using the final textual analysis workflow wizard, the gazetteer wizard, the user can associate proper nouns in the texts with places in the Locations & Objects category, essentially building a dynamic geographical index of all known places. In the economic texts from Ras Šamra, we are interested in identifying the various villages and estates throughout the kingdom that are obliged to send taxes of various sorts to the central administration. In general terms, scholars

36 The practice of adoption in the ancient world includes relationships that would be more properly described as being based on economic terms than as being based on terms of familial guardianship. These business arrangements are sometimes referred to as fictive kinship relationships (Bordreuil 1981).

have been able to place the villages into regions throughout the kingdom.[37] This has been achieved primarily through a detailed analysis of the co-occurrence of place names in royal administrative lists. It is presumed that cities that are frequently listed together are likely to have been located near each other. This conclusion could be tested in RSTI and further expanded to include other considerations that leverage the vast network of data.

With the textual data well modelled and described by properties, the user can begin to employ powerful queries to investigate the data. I will highlight two of the more interesting query types: the co-occurrence and sequence queries. These two queries are variations on a theme. The co-occurrence query allows the user to find any texts that attest a list of words, no matter the order in which they appear in the text. One may even query categories of words based on grammatical or other properties. The sequence query performs a similar search but returns only texts in which the words occur in a given order. This type of query may seem trivial on its face, but consider the complexities of ancient writing systems. How would we search for texts that contain the words for beer, distribution, and workers when each of these words may be spelled in five or six different ways? If we were dealing with textual data stored as lines in a text, this query would be very difficult. However, because textual data in OCHRE is organized into lemmata, with phonemic and attested forms contained therein, the query can look for texts that attest the lemma, regardless of the attested spelling. The query simply follows the links from the lemma in the dictionary to the discourse units in the texts.

Conclusion

Textual data is complex. As scholars who study this complexity, we need digital tools that accurately model this complexity. If we are to employ a database in our work, it must meet the challenges of the textual data. RSTI uses OCHRE for this very reason. The item-based semistructured data model suits textual data more appropriately than any other current data model. The reality is that epigraphers and philologists make observations at every level, from the entire text down to the individual sign. For this reason, it is our view that textual data should be atomized into individual database items that represent the signs of a text.

In addition to meeting these basic criteria, OCHRE also provides powerful analytical tools to guide the scholar through common analytical activities.

37 Soldt 2005.

From generating dictionary entries based on texts to documenting social-network relationships, the scholar can work with their data in a single unified platform. Data can be reformatted and exported for use in other programs, but the primary task of describing the data can take place entirely in OCHRE. The centralized database obviates the need to maintain a separate image database, GIS database, text database, and object inventory. The OCHRE Java client makes it possible for users without programming or advanced computational skills to work with their data in a natural and familiar way. Even though the underlying data is stored in XML, the user need never interact directly with these files.

While OCHRE was developed for the archaeology and complex languages of the ancient world, the underlying data model and tools are just as applicable for modern languages or archaeological sites in other contexts. Logosyllabic Akkadian is complex, so RSTI requires all the complexities offered by OCHRE. But the available complexities of the system may be used only when needed. With a highly customizable taxonomy and flexible data model, OCHRE is ready to accommodate a wide variety of projects.

References

Akkermans, Peter M. M. G., and Glenn M. Schwartz. 2003. *The Archaeology of Syria: From Complex Hunter-Gatherers to Early Urban Societies* (*c. 16,000-300 BC*). Cambridge: Cambridge University Press.

Bordreuil, Pierre. 1981. "Production-pouvoir-parenté dans le royaume d'Ougarit (14ème-13ème s. av. J.C. environs)." In *Production, pouvoir et parenté dans le monde méditerranéen de Sumer à nos jours: Actes du colloque / organisé par l'E.R.A. 357, CNRS/EHESS, Paris, décemebre 1976*, edited by Claude-H. Breteau, 117–131. Paris: Geuthner.

Bordreuil, Pierre, and Dennis Pardee. 1989. *La Trouvaille Épigraphique de l'Ougarit 1: Concordance*. Ras Shamra-Ougarit 5, no. 1. Paris: Éditions Recherche sur les Civilisations.

Bordreuil, Pierre, and Dennis Pardee. 2009. *Manual of Ugaritic*. Linguistic Studies in Ancient West Semitic 3. Winona Lake, IN: Eisenbrauns.

Borger, Rykle. 2004. *Mesopotamisches Zeichenlexikon*. AOAT 305. Münster: Ugarit-Verlag.

Dirksen, Dieter, and Gert von Bally, eds. 1997. *Optical Technologies in the Humanities: Selected Contributions to the International Conference on New Technologies in the Humanities and Fourth International Conference on Optics within Life Sciences OWLS IV Münster, Germany, 9–13 July 1996*. Series of the International Society on Optics within Life Sciences 4. Berlin: Springer.

Eisenstadt, Shmuel N., and Luis Roniger. 1984. *Patrons, Clients, and Friends: Interpersonal Relations and the Structure of Trust in Society*. Themes in the Social Sciences. Cambridge: Cambridge University Press.

Matoïan, Valérie. 2008. *Le Mobilier Du Palais Royal d'Ougarit*. Ras Shamra-Ougarit 17. Lyon: Maison de l'Orient et de la Méditerranée.

Schloen, David, and Sandra Schloen. 2014. "Beyond Gutenberg: Transcending the Document Paradigm in Digital Humanities." *DHQ* 8 (4). <http://www.digitalhumanities.org/dhq/vol/8/4/000196/000196.html>.

Soldt, Wilfred H. van. 2005. *The Topography of the City-State of Ugarit*. AOAT 324. Münster: Ugarit-Verlag.

Thuraisingham, Bhavani. 2002. *XML Databases and the Semantic Web*. Boca Raton, FL: CRC Press.

Woods, Christopher, Geoff Emberling, and Emily Teeter, eds. 2010. *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. Oriental Institute Museum Publications 32. Chicago: The Oriental Institute of the University of Chicago.

Yon, Marguerite. 2006. *The Royal City of Ugarit on the Tell of Ras Shamra*. Winona Lake, IN: Eisenbrauns.

Yon, Marguerite, and Daniel Arnaud. 2001. *Études Ougaritiques*. Ras Shamra-Ougarit 14. Paris: Éditions Recherche sur les Civilisations.