Unlocking Legacy Data: Integrating New and Old in OCHRE

Miller C. Prosser[1] *– Sandra R. Schloen*[2]

Abstract: Given the proper tools and a clear view of the goal, it is entirely possible to unlock legacy data for use in current and future research. In our work with archaeology and philology projects at the Oriental Institute of the University of Chicago, we have come to expect that database projects usually begin with a set of pre-existing legacy data from excavation paperwork, to top plans, to supervisors' notes. These data are valuable and often the basis for the project going forward. A certain amount of conversion and janitorial work is necessary to convert the data to a digital format. Each project must decide how much of the legacy data to convert and how detailed the data should be. The item-based approach of the OCHRE database allows a project to itemise their data to the finest degree needed. These highly granular data are organised by a combination of hierarchies, links and other strategies. Once digitised and properly organised, these data become part of the larger network of research data. In this context, value is added by integrating resources of various formats (GIS, photographs etc.). Once fully integrated, the data – both new and old – can be leveraged for research and presented in dynamic and interactive online formats.

Keywords: OCHRE, graph database, legacy, integration, archaeology, philology

Introduction

We find ourselves at a peculiar point in history. While clearly working in the digital era, we are still near enough to the pre-digital era that we continue to rely upon and interact with information predating the digital revolution. We are straddling that important watershed moment called the digital revolution. On our side of this historical watershed, we have at our disposal powerful computational tools to aid our research. Aerial photography, digital total stations, tablet computers and databases – all things that used to be too expensive or too technical – are all now common and widely available for use in archaeological field work. Most research now leverages at least some set of computational tools. However, most research is based also in part on information that predates the digital revolution. This remains the central issue: archaeological and philological projects often face the challenge of integrating data from a previous decade or even century. At the very least, our research would be enriched if we were able to incorporate related data from previous research projects. One may even argue that we have a responsibility to include these previous data. But legacy data typically are different than born-digital data in form and structure. As such, we are faced with the problem of how to incorporate information from the previous side of the historical watershed; in other words, an era when the use of databases was not common in the humanities and published research took the form of printed papers and volumes. For some researchers, their immediate reaction is that a current or proposed project would simply lack the time and resources to digitise and incorporate information from previous research. Even an earnest and thoughtful approach might lead one to conclude that a given set of legacy data simply reflects a different approach to describing the research, an approach that would not align neatly with the current research project. As such, even if one were to digitise the legacy data, it would still not align with the approach employed by the current project. With this realisation, a

[1] Miller C. Prosser, PhD (m-prosser@uchicago.edu) is a research database specialist at the OCHRE Data Service (ODS) of the Oriental Institute and a lecturer in the Digital Studies program at the University of Chicago.

[2] Sandra R. Schloen (sschloen@uchicago.edu) is the OCHRE developer and the manager of the ODS of the Oriental Institute of the University of Chicago.

researcher may be tempted to conclude that any attempt to integrate legacy data might derail the goals of the current project. These are common conclusions, especially in the humanities, where researchers rarely agree on descriptive terminology and certainly rarely use the same tools and methods for collecting data.

These defeatist conclusions may have been understandable in the early years of the digital revolution; however, database applications have evolved to address these precise problems. Whereas in a previous era a database may have been limited by a data model that required tables to be joined by common fields, current technological advances make it possible to use a data model that is better suited to the unique nature of humanities research data and which specifically supports the desire to integrate legacy data. With this technological hurdle no longer impeding the task of integrating legacy data, researchers now need only apply themselves to the task.[3] We hope to demonstrate that with the appropriate tools and strategies – specifically a data model that allows for the integration of disparate datasets – and with a clear view of the goal, legacy data can be unlocked and can play an active role in an ongoing research project.

Case Studies

Before discussing the data model and strategies that make this task possible, the following examples demonstrate the successful integration of legacy data by some of the projects using the Online Cultural and Historical Research Environment (OCHRE).[4] Like most research projects, these all began with various sets of legacy data – some in the form of printed materials or even in the form of handwritten records. Other projects started with spreadsheets and tables from relational databases. These projects also faced the problem of integrating images from slides, prints and digital media as well as maps and top plans, both in print form and in GIS formats generated in programs like ArcMap.

The first example, Tell al-Judaidah, is a site excavated by Robert Braidwood and the Oriental Institute in the 1930s.[5] It is one of the largest sites in the Amuq valley in southern Turkey and has a long occupational sequence from the Neolithic to the Byzantine Period. The Oriental Institute later returned to the site in 1995 after a bulldozer exposed a significant mudbrick structure.[6] In the years that followed, Dr. Lynn Swartz Dodd continued studying the small finds and organising the data of the excavations using OCHRE. She was faced with handwritten records and photographic slides from the excavations in the 1930s, to which she would need to add her own digital images and observations. The process of digitising the original handwritten documents and images was a fairly simple matter of scanning them. A simple scanned document, however, does not magically become data. Dodd and her team used tools in OCHRE to create live hot spot links that associated object photos with the handwritten documents describing where the objects were recorded and described.

Figure 1 shows a scan of the handwritten excavation record from the 1934 season. The yellow polygon areas on the left side of the scan represent live links to the images of the items in question. Item 4Z is shown in the picture, which itself is a scan of a slide produced in 1934. Professor

---

3 We do not wish to understate the resources necessary to convert, upgrade, correct, and generally curate a legacy data set. We flatly reject, however, the assertion that such a task is impossible.

4 OCHRE is a computational platform, supported by the ODS of the University of Chicago, and available to academic research projects everywhere. The OCHRE application is free to download (cf. OCHRE Data Service). Projects register with the ODS for data hosting, technical consulting and support, and legacy data conversion services. To read more about the ODS and how to start an OCHRE project, see <https://ochre.uchicago.edu/> (last accessed 18 Dec. 2019).

5 Cf. Braidwood – Braidwood 1960.

6 Cf. Yener – Wilkinson 1996.

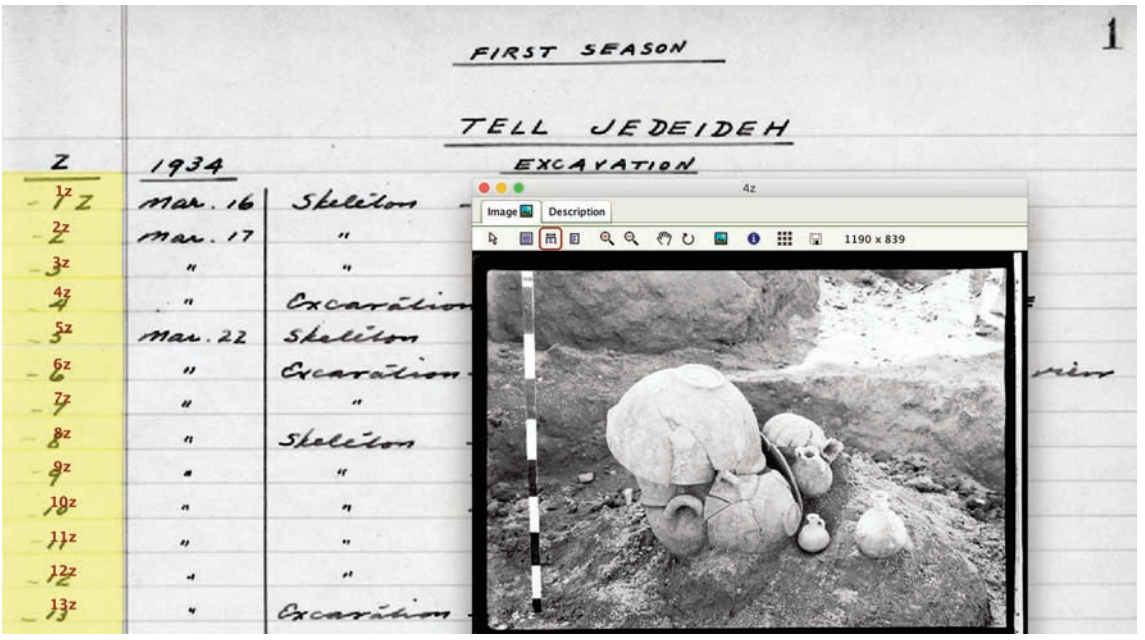

Fig. 1 Tell al-Judaidah scanned document and object photograph in OCHRE (courtesy of the Tell al-Judaidah project of the Oriental Institute)

Dodd and her research team continue to curate the legacy data, adding their own analysis and photographs.

One of the more extensive case studies in which the ODS has participated is the Computational Research on the Ancient Near East (CRANE) research project. This project seeks to provide a framework for studying archaeological data from various sites throughout the Orontes watershed region of Syria and Turkey. At the project's inception, team members assembled research data from excavation and survey projects including the Homs Regional Project, the excavations at Tell Acharneh, the Neubauer expedition to Zincirli and the excavations at Tel Tayinat.[7] Each project brought with it a variety of legacy data, including digital data recorded in spreadsheets and relational database programs. The initial task for this project was to export the data from the old relational databases and prepare it for import into OCHRE. As we describe below, OCHRE allows each project to retain its own nomenclature. In other words, none of the projects were forced to change their data to match the recording methods of another project. The end result is data from various projects that can be queried and viewed together. For example, as of the current writing, a simple query for 'jewellery' across all CRANE projects returns 302 results. If we add 'material' as a search criterion, we can search for metal jewellery, a search that narrows the results to 102. We can search more specifically for metal jewellery which is further described as gold. A search for 'gold jewellery' returns 8 results, a mixture of pendants and bracelets from different sites. Had the various sets of legacy data from these projects not been centralised in OCHRE, this sort of query would have been extremely difficult.

In addition to data related to excavation details and small finds, the CRANE sub-projects each have a significant collection of GIS data. Typically, this type of legacy data would remain isolated and available only in the GIS program in which it was created. However, OCHRE provides a method for integrating all GIS data into the project database where it can be used along with other project data. Each of the separate sub-projects has its own set of top plans and shapefiles,

[7] The Homs Regional project is directed by Graham Philip (Durham University). The excavations at Tell Acharneh were carried out under the direction of Michel Fortin and Elisabeth Cooper (Université Laval). The Neubauer expedition to Zincirli is co-directed by David Schloen (University of Chicago) and Virginia Herrmann (Universität Tübingen). The Tayinat Archaeological Project is directed by Timothy Harrison (University of Toronto).

but the unifying CRANE parent project, which hierarchically contains the individual site projects, also includes a set of GIS data that will play an important role in the modelling of ancient climate data, which pertains to the study as a whole. Researchers are currently gathering climatological data for the Orontes watershed area of interest. This GIS data will serve as one component used in addressing the question of human impact on the environment through agricultural and pastoral subsistence practices. Information such as soil types, land cover by tree species and vegetation, although created in ArcMap and stored as GIS shapefiles, can be used as part of a broader investigation that includes core project data in OCHRE. In short, OCHRE reads and integrates GIS data produced in standard programs like ArcMap or QGIS. This tight integration of OCHRE data and external geodata means that a project can continue using the features of their preferred GIS program, while also gaining the advantage of integrating GIS data with the rest of their data.

The third and final case study comes from the Ras Shamra Tablet Inventory, a project co-directed by Miller Prosser and Professor Dennis Pardee. This project began with a single Microsoft Word document recording the find spots of all the inscribed objects from Ras Shamra–Ugarit.[8] This simple list of thousands of objects and their find spots was converted into a hierarchical system of organisation representing the spatial relationships among the excavation areas and the archaeological finds at the site of Ras Shamra. Each inscribed object finds its home in the database at the lowest level of the hierarchical path which represents the most specific context known for that object. For example, RS 6.021, an inscribed stele, is found in the following hierarchical organisation.

The Site of Ras Shamra
> The Acropolis
> Temple of Dagan
> Topographic Point 715
> RS 6.021

This object is situated in the broader scope of the site, inheriting its context from its place within the hierarchy. The project is currently integrating legacy data from printed publications, including architectural plans and other maps that specify find spots. These printed maps are scanned and georeferenced in ArcMap, then added as resources in OCHRE.

Figure 2 displays the data about the stele and shows its find spot on the georeferenced archaeological top plan.[9] Note that OCHRE makes every effort to respect the spirit of legacy data, not requiring more than can be accurately stated, nor requiring that it conform to modern standards. An older standard of recording might have indicated that the stele was ‘outside the main entrance to the Temple of Dagan’. Today we would use a high resolution instrument to capture exact coordinates at a high degree of precision. OCHRE provides sufficient flexibility and a variety of recording methods so that legacy data can be represented appropriately.

The more primary task of the Ras Shamra Tablet Inventory is the creation of reliable text editions, including text transliterations, commentary and object photographs where available. These various datasets come from many sources: text editions from Pardee, digital images produced by members of the Mission de Ras Shamra and a variety of printed sources. The result is an interactive text edition that includes epigraphic commentary, lexicographic analysis, translation and images.

Figure 3 shows a view of an administrative text written in alphabetic Ugaritic (RS 15.022+). The digital image includes outlined letters, each of which links to the respective letter in the transliteration.

8 Cf. Bordreuil – Pardee 1989.
9 Callot 2001, 167, fig. 44.

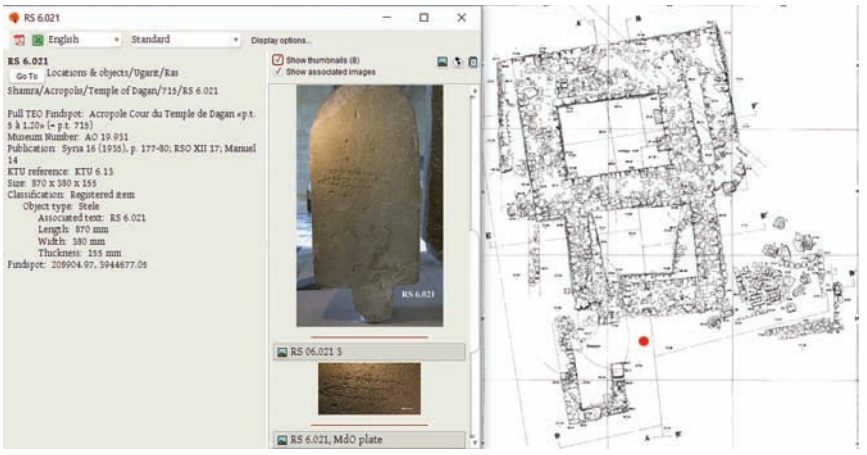

Fig. 2 Ras Shamra Tablet Inventory, RS 6.021 find spot in OCHRE
(© The Ras Shamra Tablet Inventory; image © The Mission de Ras Shamra/PhoTÉO)

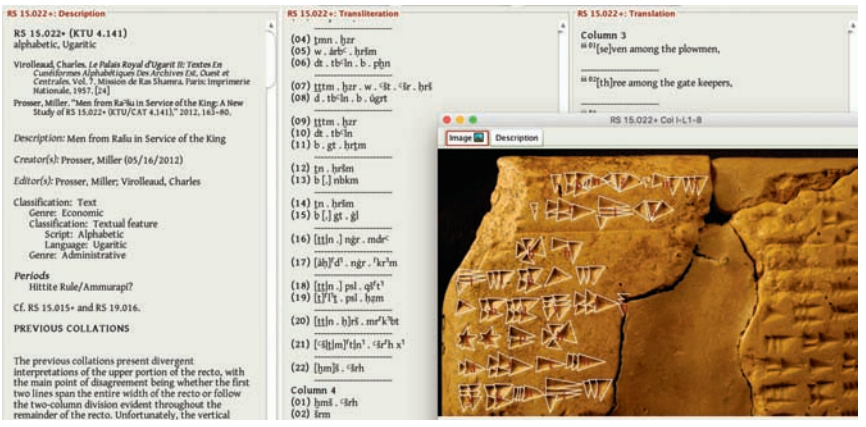

Fig. 3 Ras Shamra Tablet Inventory, RS 15.022+ in OCHRE
(© The Ras Shamra Tablet Inventory; image © The Mission de Ras Shamra/PhoTÉO)

OCHRE Data Service

The current authors are core members of the OCHRE Data Service (ODS), a team of scholars and technical experts who support various research projects using the OCHRE platform. In the realm of digital humanities, we are equal parts *digital* and *humanities*. Schloen, a computer scientist, is the OCHRE developer. Prosser earned his doctorate in Northwest Semitic Philology at the University of Chicago. This combination allows us to understand the research questions asked by our research colleagues and apply computational solutions not typically available to scholars. Researchers using OCHRE work in a broad range of topics and locations, from archaeological field work in Israel, Syria, Turkey and Niger to detailed philology projects involving languages such as Akkadian, Aramaic, Egyptian (hieroglyphic and Demotic), Elamite, Hittite, Old Assyrian and Ugaritic. Our users are located all over the world, from Los Angeles to Toronto and from Be'er Sheva to Tübingen.

Because we work within an institution and a field of study with a long and rich history, we have seen every imaginable type of legacy dataset. One of our frequent and primary tasks is to help researchers integrate legacy data into their active, digitally based research project. As we consult with researchers and advise on computational strategies for modelling research data, we apply lessons we have learned over the years. Because we deal with legacy data on a regular basis, we can save a research project time by avoiding common pitfalls. Whether it is something as simple as the proper strategy for naming digital images and organising them on a remote server

or something more difficult like strategies for recording uncertainty and disagreements in the data, ODS can provide recommendations for best practices. Certain strategic missteps early in the process can make the integration of legacy data more difficult than necessary, but with some guidance and the right tools, the once impossible task becomes possible. The OCHRE database environment is uniquely suited to this task.

Initially conceived in 1989 for the purpose of gathering data to do prosopographical analysis of the personal names mentioned on the Ras Shamra tablet inscriptions, OCHRE has grown to accommodate many types of data and support many different research goals. Whether representing the cuneiform signs on these ancient tablets from Ras Shamra, the typeset manuscript of a Shakespeare folio or the inked cursive of the letters of Charles Darwin to his contemporaries; whether describing the paleoclimate of the Orontes watershed region in south–central Turkey or last season's excavation at Tel Keisan in Israel; whether collecting inscriptions on stones in medieval South Indian temples or inscribed coins from ancient Greece, OCHRE's flexible, item-based data model provides structures without strictures. Data, both legacy and born-digital, from over 50 active projects to date, representing over 8.5 million database items, are integrated and managed by OCHRE in a native XML database[10] secured and supported by the Digital Library Development Center at the University of Chicago.[11] From the start, the OCHRE system was designed to handle both variety and volume of data, and its many successful use cases prove the model.

Dealing with Large Data Sets

At times we may find the prospect of dealing with legacy data too daunting because the size of the legacy data is so vast. No doubt, a mountain of data can be intimidating, but with the appropriate tools for cleaning up and systematising the data, the task is not so different, whether the dataset includes 100 or 15,000 items. Free and simple to use tools such as Advanced Renamer and OpenRefine make the data janitorial task easier than it would be otherwise.[12] The former can rename thousands of files at once, removing characters that are typically not valid in file names or imposing order to make files more sensibly sortable. The latter is a powerful tool that helps the user refine spreadsheets with thousands of rows of data. Its built-in features allow quick and easy correction.

When free tools are not enough to solve all issues in a legacy dataset, ODS has the experience and some customised tools to help in the janitorial process. Especially for philology projects, whose legacy data may predate the Unicode standard, ODS has created custom upgrade utilities to convert old documents to current formats and meet the Unicode standard. Most of the writing systems of the ancient world have been added to the Unicode standard, but some only in recent years, which leaves some of our colleagues with documents that were created with customised non-Unicode fonts. In a recent effort to solve this problem at the Oriental Institute, ODS converted tens of thousands of old Microsoft Word documents to Unicode. Documents containing text transliterations can then be imported into OCHRE with the knowledge that the character encoding and display will be the same for all viewers.

The OCHRE database environment has no limit to the number of legacy items it can accept. Flexible import and synchronisation tools allow legacy datasets to be ingested into current projects. We have added thousands of images at once, automatically linking them to other database items where possible. We have imported tens of thousands of faunal remains items from old

[10] Tamino, from Software AG.

[11] DLDC.

[12] Advanced Renamer is freeware developed by Kim Jensen. OpenRefine (formerly a Google product) is an open source and free tool.

spreadsheets. Through this semi-automated import process, these faunal remains were converted from a single spreadsheet into tens of thousands of database items, ready to be viewed, queried and analysed in various ways.

OCHRE

Without attempting to present a complete history and survey of the evolution of databases, we will observe simply that it is important to recognise that there is no single approach to modelling data in a database. Of the various ways to structure data, most are familiar with the relational data model.[13] This is the underlying structure of many database platforms made for the business world. In this model, data are recorded in a series of tables which are related by key fields. The related fields define how information in one table is associated with information in other tables. The related field is typically something that occurs in all related tables, like a customer number, an object ID or some other common datum. With the proper configuration, this data model can be made to work for archaeological data; however, it is inherently limiting and difficult to modify as research progresses. Tables are well-suited for recording highly regular data like invoice details or names and addresses, but they are not the best structure for recording highly irregular data such as faunal remains, pottery classification or text transliteration. When applied to these types of data, the table will either include a cumbersome number of columns, each of which records a sparse amount of data, or a limited number of columns, each of which contains a cumbersome amount of data. In our experience, this data model is difficult to apply to humanities research data.

Itemisation and the Item-Based Approach

The OCHRE data model uses an item-based approach to organising data. Instead of recording data in tables, OCHRE records each unit of observation as a discrete database item. Data is not stored in tables, but this need not induce panic. Items can be reorganised, combined, collected and represented in tables for ease of viewing, analysis or exporting. The important distinction is that in the OCHRE data model, a table is a derived data structure instead of the primary data structure.

Because the item-based data model is not as commonly known, it requires some introduction. Imagine if you could treat each of your artefacts, faunal remains, or botanical samples as its own thing. Each object the archaeologist finds, sign the philologist reads or image the photographer produces – each stands as its own discrete thing, a database item. Instead of grouping items together in categories, some of which are artificial, a database item can be recorded independently of any other thing. One does not excavate a collection of like items from the ground. An archaeologist always hopes to stumble upon a cache of objects, of course! Even so, items are typically excavated one at a time. The item-based data model conveniently parallels this reality. In the real world, things exist as individual items. This simple innovation, that every thing is its own item, is the basic premise of the item-based data model.[14] This approach often feels very natural to the archaeologist who is trained in dealing with individual artefacts.[15] To be clear, in this data model an object such as a pottery vessel, tablet, bone or any other small find is what we call an item.

---

13 The foundational work on the relational model is Codd 1970.

14 From a technical perspective, the implementation of an item-based data model is better suited to what is known as a graph database rather than a relational database.

15 In cases in which the archaeologist wishes to record a collection of items without itemising each item, OCHRE can treat the collection as an item. For example, it is common to treat a set of non-diagnostic pottery sherds as a collection. They are important in their collective count and weight, and a researcher will never comment specifically on any single sherd.

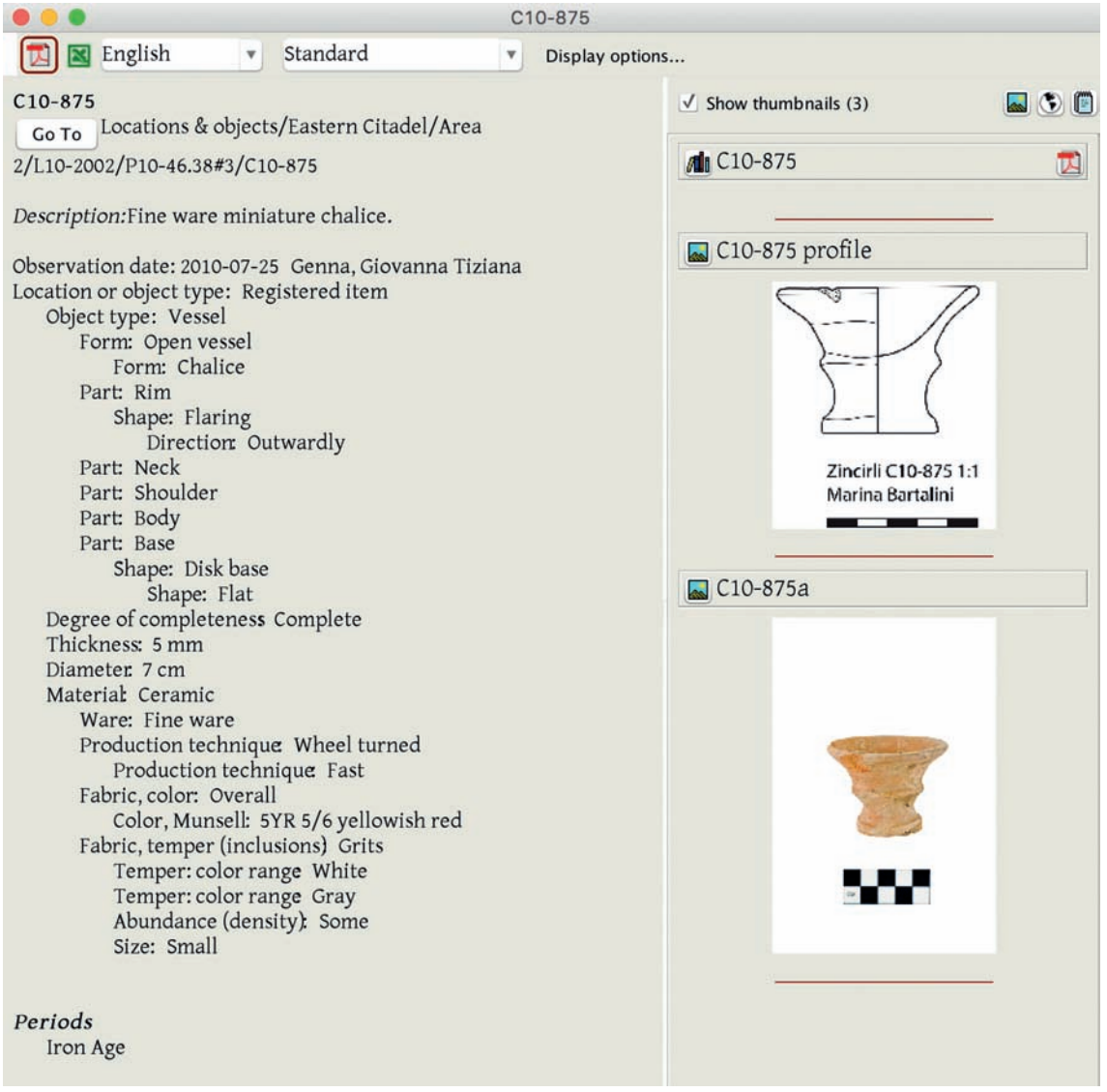

Fig. 4 The Chicago-Tübingen Archaeological Project in Sam'al, the OCHRE entry for vessel C10-875 (© The Chicago-Tübingen Archaeological Project in Sam'al)

The word 'item' also refers to various other data points such as architectural buildings and rooms, modern and ancient people and even periods of time. All of these things are distinct database items. The item-based approach simply refers to the strategy of recording information as distinct items. The process of itemisation may be characterised by batch conversions of spreadsheet tables into many discrete items, dividing larger blocks of data into discrete units or simply the task of entering new information as items from the start. At first reading, the process may sound overly technical, but it couldn't be simpler. Any given thing, a minimal meaningful unit of observation, is an item.

Even PDFs and image resources are database items. In the case of PDFs or other documents, the project is free to decide the degree to which the text of the document is itemised. For example, some archaeological projects decide to keep intact the daily journals produced by area supervisors; they will link a digital copy of the PDF to the database item representing the relevant area. Another project might divide the paperwork into daily journals by excavated locus and link the smaller PDFs to database items representing the relevant loci. While further atomising or itemising the document into smaller units may provide more flexibility for linking the document in highly surgical ways, it is often impractical to do so. In cases such as this, OCHRE provides a

method for referencing the specific page of a PDF in a discrete database link so that the user has the flexibility of referring to specific parts of a larger document without having to overextend the idea of itemisation.

In the OCHRE approach, every database item can be identified and described by properties. Ceramic vessels frequently are described by their fabric characteristics, vessel type, decoration and other qualities. Every item can be described with properties that apply only to that item and not to an entire table or class of items. Properties themselves are database items that are applied individually and only as necessary. If a ceramic vessel has a specific decoration, then the properties required to describe that decoration can be applied. If it is a plain vessel, then these individual properties do not apply. To be clear, it is not that the property describing the decoration is left blank for a plain vessel, but that it is not even present when unnecessary. Figure 4 shows the properties for a specific vessel from Zincirli. Note that there are no blank fields. The observer recorded only the properties that applied to this specific item.

One of the greatest challenges to integrating datasets with heterogeneous origins is the typical misalignment of organisational systems. One project may have grouped all jewellery in one table. Another project may have grouped jewellery together with other decorative items. While it might be difficult to align data from two projects using a relational data model, it is much more difficult to integrate data from multiple projects when some of them are highly detailed site-based excavations and others are more cursory surveys. As described briefly above, this is precisely the type of legacy data that has been integrated in the CRANE project. In short, the item-based approach makes this possible because it unlocks the data from the restrictions of the table structure. Once each data point stands as a discrete database item, it can be properly organised and related to any other database items. Within the CRANE project, for example, any small find can be defined with properties that allow the item to be related to all like items, whether the item was found deep in a modern excavation context, lying on the top of an ancient tell or in the handwritten record of an historical field journal.

Organising Items

What do we do with all these items, if not organise them in tables of like items? As already hinted above, OCHRE organises millions of database items in hierarchies. A hierarchy is simply a tree-like structure for recording data. At each level of the hierarchy, the data may branch into any number of sub-branches. On a very basic level, each child item in the hierarchy has only one immediate parent, but a parent may have an unlimited number of child items. This strategy presents a potential limitation: how do we record the location of an item that appears in two conceptually distinct hierarchies? To overcome this apparent limitation, OCHRE allows any given database item to be contextualised in multiple hierarchies. A database item is not limited to a single hierarchical context. It can be reused in as many contexts as needed. In practise, this means that any given locus of excavation for an archaeological project can be recorded in a hierarchy that represents a configuration of excavation units like grids and squares, but it can also be contextualised in a separate spatial hierarchy that represents the ancient architectural system of neighbourhoods, buildings and rooms. In what can be called a polyhierarchical strategy, the locus exists as a single item with a unique identifying number but is contextualised within two hierarchies. The important point here is that the locus is not represented by two separate database items, but rather a single database item organised in two hierarchies. In Figure 5, we see two spatial hierarchies from the Neubauer expedition to Zincirli. Notice that L08-5019 appears as part of a room (on the left) and in its original excavation context (on the right).[16] This database item happens to represent a wall.

[16] The Zincirli nomenclature system uses L in an item name to represent a locus. Other projects use their own terminology such as lot or level or unit.

Fig. 5 The Chicago-Tübingen Archaeological Project in Sam'al, a wall (L08-5019) in two hierarchical contexts (© The Chicago-Tübingen Archaeological Project in Sam'al)

When it was initially excavated, the archaeologists could not yet determine its place within the ancient architecture, so it was recorded in an excavation context. After further excavation, the same database item representing this wall was organised secondarily within a hierarchical structure of ancient buildings and rooms.

A hierarchy can represent spatial organisation from region to site, area, grid, square, locus or finegrid, for example. Each level of the hierarchy is contained by its parent item and contains its child item(s). Recalling the example above of the stele from Ras Shamra, the database item that represents the site of Ras Shamra is the parent item to all of the excavation areas on the site. The acropolis of the site, one of the child items of the Ras Shamra item, is also a parent item to more specific areas, each of which is the parent to the various topographic points, which are in turn the parent items to all of the registered items found at each point. This system of hierarchical containment proves to be a very powerful and highly flexible approach to organising spatial data.

Time periods can also be represented in a hierarchical organisation. A sequence of historical periods or archaeological phases when organised into a hierarchy defines the relationship between each nested level of the hierarchy. The order of the periods in the hierarchy defines the chronological order of the periods or phases. Also, if one period in the hierarchy is the parent of other periods, this indicates that the parent item is a broader period that is further divided into sub-periods represented by the child items. Iron Age IIA and IIB are both periods contained by the higher level item Iron Age II, which is contained by the higher level item Iron Age. On a practical level, this allows the user to specify the period of an item without having to over-specify. A given small find, for example, may be identifiable only generally as dating to the Iron Age. Another might

be tagged more specifically to the Iron Age IIA period. The hierarchical arrangement of periods allows the user to query for a specific sub-period or a broad period. A query for all items from the Iron Age would find items identified at the highest level as well as all items tagged specifically to any of the periods contained within the Iron Age. In contrast, a query for items dating to Iron Age IIA would find only items from that specific period, excluding any items identified only generally as dating to the Iron Age. A project can define as many period hierarchies as necessary. Typically, archaeological projects include a general outline of historical periods, but may also include regional and site-specific phasing systems or even political outlines. This flexibility is particularly important when capturing legacy data whose details were not captured to the modern scientific standard but are broader or vaguer, yet no less important. We capture what we can without needing to over-specify and without needing to conform to current expectations.

Hierarchical organisation provides a limitless and flexible solution for organising database items regardless of their source, whether born-digital or non-digital and current or legacy. Because the structure of a hierarchical organisation can be revised easily to reflect new and updated understandings of the data, it is uniquely appropriate for archaeological research. New loci and architectural features can be added to existing structures, or new arrangements of existing hierarchies can redefine the related database items. Legacy data can be added to existing structures or can supplement existing data to add missing branches. For example, site-based excavations like Tell Keisan and the Jaffa Cultural Heritage Project have integrated the excavation data from areas exposed by previous teams.[17] Where these areas overlap with the current expedition, the two can exist in the same structure. Otherwise, the new and legacy areas may exist as separate sibling items in the broader hierarchical structure that represents the entire site.

Taxonomy and Thesaurus

In OCHRE, variable-value pairs defined in a project taxonomy are created to represent the descriptive properties of database items. A new project may choose to adopt variables and values established by other projects, but they are also free to customise their project taxonomy to fit their own project needs or accommodate terms from legacy data. Again, hierarchical structure is used to organise the taxonomic values, which themselves are also database items. The taxonomy variable ‘Vessel type’ is the parent of various child values such as Bowl, Juglet, etc. The taxonomy is flexible both in its organisation and the specific terminology used.

We have argued that the item-based approach solves the problem of the misalignment of legacy datasets by freeing data from the strictures of tables, but what about cases when the legacy dataset uses a completely different nomenclature than the current project? As is well known, there is a great deal of variation in the field of archaeology. In fact, we may have already used certain terms that are foreign to the reader. Some projects do not use the term locus, or terms for an excavated unit of soil may be bucket, pail or some term derived from a local dialect (e.g., goufa). In the case where a legacy dataset uses different nomenclature, OCHRE provides a mechanism for creating a thesaurus of terms. Any given variable or value in the taxonomy can be equated with another item in the taxonomy. If one site uses ‘area’ and another uses ‘field’, neither project need adopt the other’s nomenclature. In an ideal world, one might hope that common standards might be agreed upon, but in the real world this seems unlikely. Even so, we would still be faced with legacy data that was collected before such an ideal intellectual détente had been achieved.

[17] The current excavation at Tell Keisan is directed by David Schloen (University of Chicago) and Gunnar Lehmann (Ben-Gurion University). The Jaffa Cultural Heritage Project is co-directed by Aaron Burke (UCLA) and Martin Peilstöcker (Johannes-Gutenberg Universität, Mainz, Germany). Cf. JCHP.

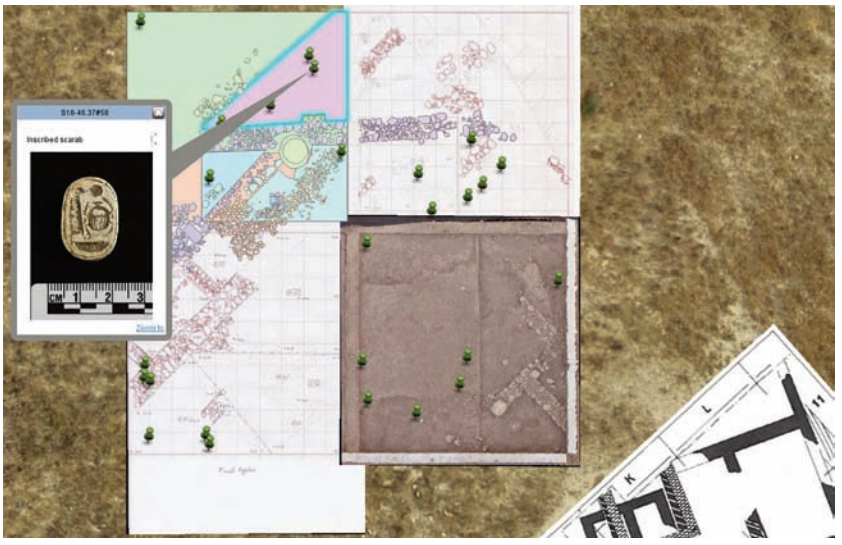

Fig. 6 Tell Keisan GIS view in OCHRE (© Tell Keisan Excavations)

Data of Different Types

OCHRE is a comprehensive data management system. Regardless of the type of data, the goal is to integrate all data in one place. This applies to digital images, PDFs, shapefiles, raster images, drawings, 3D imagery and files of any variety of other formats. The Ras Shamra Tablet Inventory has acquired and integrated legacy datasets from several different sources. The first major set of digital images included the thousands of digital images produced by the epigraphic team of the Mission de Ras Shamra during various research trips to the National Museums of Damascus, Aleppo and the Louvre.[18] Another set of legacy data comes from digital scans of maps and top plans published by the Mission. These were georeferenced in ArcMap and added to the project. From the recently acquired set of research photographs produced by John Ellison during his dissertation research in Damascus and Aleppo, over 10,000 images in all were imported and linked to the database items representing the tablets pictured.

On another front, the CRANE project is currently upgrading and integrating GIS data collected over the years in various formats. In a way, the project created its own set of outdated data simply by using a wide range of GIS collection methods and utilising different technologies as they improved over time. Not all legacy data comes from outside sources as it turns out! Site maps, top plans and shapefiles, georeferenced to the spatial framework of the project's excavation area, can be catalogued in OCHRE and viewed alongside core project data using OCHRE's tightly integrated GIS features. With the right tools in place, the project's geospatial data can contribute to the ongoing picture of the excavation.

One of our core principles of data management is to reduce the number of independent datasets, or inversely to integrate all data sets. OCHRE thus serves as a repository for all of a project's data. This is true as much for legacy data as for all other types of data. In the specific case of GIS data, the goal is that any item in the OCHRE database with spatial information can place itself on a map. Figure 6 below illustrates a view from the Tell Keisan project with selected architectural features drawn in ArcMap. Find spots of small finds, shown as green pushpins, which are live links to object photographs, scanned top plans and part of a scanned map from the previous

[18] Photography credit goes to Dennis Pardee, Robert Hawley, Carole Roche-Hawley and Miller Prosser, with image copyright belonging the Mission de Ras Shamra/PhoTÉO.

expedition to the site are visible. This single view shows the complete integration of various data types, digital and non-digital as well as newly collected and legacy.

Conclusion

A clear vision of the desired outcome, a team of determined and hardworking colleagues and the right computational tools are the components that make it possible to integrate legacy data with current project data. The ODS works with its research partners to define the desired outcome and directs project personnel on any janitorial work that may be required to upgrade legacy data. In terms of computational approaches, experience has taught us that OCHRE's item-based data model is more flexible and powerful than the relational data model, especially for dealing with the unpredictable and heterogeneous data that archaeologists and philologists create. In the end, it is satisfying and rewarding when a project unlocks the value of its legacy data and can then begin to incorporate this information into its continuing research.

References

Advanced Renamer
<https://www.advancedrenamer.com/> (last accessed 18 Dec. 2019).

Bordreuil – Pardee 1989
P. Bordreuil – D. Pardee, La Trouvaille Épigraphique de l'Ougarit 1. Concordance, Ras Shamra-Ougarit V, 1 (Paris 1989). Online <https://www.mission-ougarit.fr/wp-content/pdf/mission_ougarit_publication_rso_0510.pdf> (last accessed 18 Dec. 2019).

Braidwood – Braidwood 1960
R. Braidwood – L. Braidwood, Excavations in the Plain of Antioch I. the Earlier Assemblages Phases A–J, Oriental Institute Publications 61 (Chicago 1960). Online <https://oi.uchicago.edu/research/publications/oip/oip-61-excavations-plain-antioch-i-earlier-assemblages-phases-j> (last accessed 18 Dec. 2019).

Codd 1970
E. F. Codd, A relational model of data for large shared data banks, Communications of the ACM 13, 6, 1970, 377–387. doi:10.1145/362384.362685

Callot 2011
O. Callot, Les Sanctuaires de l'Acropole d'Ougarit. Les Temples de Baal et de Dagan, Ras Shamra-Ougarit 19 (Paris 2011).

DLDC
Digital Library Development Center
<https://dldc.lib.uchicago.edu/> (last accessed 18 Dec. 2019).

OCHRE Data Service
Online Cultural and Historical Research Environment Data Service
<http://ochre.uchicago.edu/> (last accessed 18 Dec. 2019).

OpenRefine
<http://openrefine.org/> (last accessed 18 Dec. 2019).

Software AG
<https://www.softwareag.com/corporate/products/webmethods_integration/default> (last accessed 18 Dec. 2019).

JCHP
The Jaffa Cultural Heritage Project
<http://jaffa.nelc.ucla.edu/> (last accessed 18 Dec. 2019).

Yener – Wilkinson 1996
A. Yener – T. Wilkinson, Amuq Valley Projects, in: W. Sumner (ed.), The Oriental Institute 1995–1996. Annual Report (Chicago 1996) 11–21. Online <https://oi.uchicago.edu/about/annual-reports/oriental-institute-1995-1996-annual-report> (last accessed 18 Dec. 2019).