The aim of our project, Last House on the Hill (LHotH), is to holistically reconstitute the rich multimedia and primary research data with the impressive texts of the monograph, the printed final report of the Berkeley Archaeologists at Çatalhöyük (BACH) project, in which a team from UC Berkeley excavated a group of Neolithic 9000-year old buildings at this famous cultural heritage location in Central Anatolia, Turkey. The Last House on the Hill brings together the published text, complete project database (including all media formats such as photographs, videos, maps, line drawings), related websites, data and media outside the direct domain of the BACH project, and recontextualised presentations of the data as remixes, movies, and other interpretive works by BACH team members and many others. We are achieving this through an event-centered, CIDOC-CRM compatible implementation ontology, expressed through an open source web-publishing platform, providing open access, transparency and open-endedness to what is normally the closed and final process of monograph publication.

The idea of embedding, interweaving, entangling and otherwise linking the data and media from archaeological excavations with their interpretation and meaningful presentation in an open access sharable platform has long been an ambition of those of us working in the digital documentation of archaeological research and the public presentation of cultural heritage. Formidable barriers still exist to making it possible for projects to achieve these aims, ranging from intellectual property concerns to providing commitments to the long-term sustainability of the digital content. We believe that our event-centered implementation ontology will make it far easier for archaeologists and researchers in other disciplines to organize, manage and share their data while gaining the significant benefits of the CIDOC-CRM framework.

This paper describes the strategy, goals, architecture and implementation for the project, emphasizing the novel and innovative approaches that were required to make the project successful.

Categories and Subject Descriptors: Publishing, metadata, semantics, long-term preservation and access.

Additional Key Words and Phrases: Archaeology, digital publishing, Neolithic, Anatolia, architecture, open knowledge.

ACM File Format:

1. INTRODUCTION

During the summers of 1997-2005, a team from the University of California at Berkeley (BACH team) carried out an archaeological project of excavation and analysis at the site of Çatalhöyük in Central Turkey, a 9000-year old Neolithic settlement mound,
as part of the overall Çatalhöyük Research Project. This article is about the different ways in which we share the process of work and the results of our research with the public - in other words publication - and the effectiveness of different publication formats on the sustainability and long-term accessibility, usefulness, and meaningfulness of the data that we spend our lives obtaining.

In a commencement speech at Arizona State (ASU), U.S. President Obama reminds the 2009 graduating class that life is not measured only in accomplishments, but by perpetuating actions: "That is what building a body of work is all about - it's about the daily labor, the many individual acts, the choices large and small that add up to a lasting legacy. It's about not being satisfied with the latest achievement, the latest gold star - because one thing I know about a body of work is that it's never finished. It's cumulative; it deepens and expands with each day that you give your best, and give back, and contribute to the life of this nation. You may have set-backs, and you may have failures, but you're not done - not by a long shot." [Zim09].

Archaeological fieldwork produces a fragile legacy, requiring remarkable attention and effort in order to assure that our many individual acts will be sustained into deep time. Access to our legacies, analog or digital, are not assured, not by a long shot, so long as the de facto preservation standard through publication comprises of only the synthetic accounts of these events in the form of monographs. The sharing and the preservation of human traces digitally, coherent access to these traces for future generations [Ash09], depends on a coherent reckoning of all of the evidence streams from our practice, not just the hand-picked pieces chosen for their aesthetics or relevance to the particular assertions we make in the authoritative texts of the final field reports. As challenging as it continues to be, our discipline requires us to not be satisfied with a definitive accounting as told by a select few, rather to present to our current and future audiences a full, unmitigated documentary of the choices, both large and small, that led to our conclusions about the archaeological past.

It became more and more frequent in the late 1990s and 2000s to supplement the written text of a scientific published report or a textbook with a CD-ROM or (later) DVD-ROM that was slipped into an envelope inside the back cover of the volume. Such CD-ROMs accompany most of the preceding volumes of the Çatalhöyük Research Project, specifically those reports of the 1995-1999 excavations of the 9000-year old Neolithic settlement mound in Turkey. They contain supplemental tables, texts, and figures, as well as videos that would either have been left out of the printed version because of space or format limitations (e.g., color photos, videos). The supplemental CD-ROMs were useful because they meant that archaeological publication could transcend publishers' restrictions on details of data and media. But they have proved to be a librarian's nightmare, since they are difficult to catalog and very easily go missing.

John Wilbanks, Executive Director of the Science Commons, notes "the irony that right at the historical moment when we have the technologies to permit worldwide availability and distributed process of scientific data, broadening collaboration and accelerating the pace and depth of discovery…we are busy locking up that data and preventing the use of correspondingly advanced technologies on knowledge." [Wil07]. There are many arguments to be made for keeping data locked up and managed, but we are convinced that if at all possible, the future for archaeological data will be brighter if projects embrace open data formats, licensing and standards. IOSA.it supports and promotes the dissemination and use of open archaeological data, following the Open Knowledge Definition, "a piece of knowledge is open if you are free to use, reuse, and
redistribute it. [OKF08]" It is in this spirit that we created the Last House on the Hill project.

2. THE REMEDIATED PAPER VOLUME AND ITS DIGITAL MIRROR

The UCLA Cotsen Institute of Archaeology, who will be publishing our printed report of the Berkeley Archaeologists at Çatalhöyük (BACH), entitled House Lives [TS], have decided to give up on supplemental media (CD-ROMs and DVD-ROMs) in favour of on-line digital versions of their publications. We, as archaeologists who have long been involved in digital documentation and publication of archaeology and cultural heritage, were delighted with this news. We felt that these external media had a limited lifespan and utility, not only because they could easily go missing, but also because of their regular need for migration due to physical degradation and the inevitable (unless carefully archived) obsolescence of their software. We were also aware of the disadvantages of read-only CD-ROM media, which, as with the printed word, offer a definitive closed narrative. By contrast, we feel that narratives about archaeology, history, cultural heritage, and the past - including those about the data themselves - should be anything but closed, but should always be open for expansion, critique, and modification.

The online digital mirror of House Lives that is presented in this paper - entitled the Last House on the Hill (LHotH) project - goes much further than bringing together supplemental materials and digital versions of the published texts. Its ambition, one which we have long wished to satisfy, is to embed, interweave, entangle and otherwise link the data and media from the archaeological excavations with their interpretation and meaningful presentation in an open access, sharable platform [WT00], [JT07], [TA01], [Tri04]. The project brings together the published text, complete project database (including all media formats such as photographs, videos, maps, line drawings), related data and media outside the direct domain of the BACH project, with recontextualised presentations of the data. We are achieving this through an event-centered, CIDOC-CRM compatible implementation ontology (see sections 5-6), expressed through an open access web-publishing platform. We wish to provide access, transparency and open-endedness to what is normally the closed and final process of monograph publication.

The architecture and content management practices in which the Last House on the Hill have been constructed will act, we think, as a model and an encouragement for our archaeological colleagues to share their work with the public for the long-term. Our attitude to sharing our knowledge with the public in which we make the process of our archaeological interpretation transparent in order to engage them more intensively in our work, and our attitude to breaking the strict bondage of the empirical data is, we feel, very close to that of the Çatalhöyük team as a whole, who have made all of their data and media accessible through Creative Commons licensing (see section 9). The digital expression of our work is able to be richer, more colorful, more accessible and more engaging than a black-and-white publication, and certainly represents a more intricate and entangled expression of what we do and how we think. The narratives about the BACH excavation have been built out of the rich body of data and media that are available and accessible in this digital on-line version. It is an open-ended data stream that can grow and – as long as it is well curated – can live for many decades.

The paper volume, House Lives, is the seventh major publication by the Çatalhöyük Research Project. It comprises a record of the BACH project from 1997 to 2003. The first volume of the Çatalhöyük Research Project related to the initial phase of surface work
1993-95 [Hod97]. The second volume focused on issues of the reflexive methodology [Hod00]. The third, fourth, fifth and sixth volumes present the results of excavations of three areas at Çatalhöyük, known as SOUTH, NORTH and KOPAL, between 1995 and 1999, and their analysis and interpretation [Hod05a], [Hod05b], [Hod05c], [Hod07]. Our monograph House Lives presents the excavation results of the BACH project and their analysis and interpretation in one volume. Twenty-three authors contributed to the volume, many of which were participants in the BACH project for all of its seven years.

As with the other Çatalhöyük volumes, the results of the excavations (chapters 1-5) are followed by chapters that present the analysis of excavated data (chapters 6-19), which are followed by chapters of synthesis and interpretation (chapters 20-22). However, in keeping with the reflexive methodology, we have attempted throughout the volume to avoid the separation of the presentation of data from their interpretation. There are just two color photographs in the volume. The rest are black and white versions of originally color photos. There will probably not be an index. Appendices of lists and detailed data are not included in the printed volume. Nor are complex tables and figures such as the Harris matrices that document the stratigraphic position of each of the over 2000 excavation units. It is these limitations, resulting from prohibitive costs and production in a text media format, that we have overcome through the LHotH project.

3. LINKING UP WORLDS OF DATA

Our intention for the digital version of the volume is to reconstitute holistically the rich multimedia and primary research data with the impressive texts of the monograph. The requirements for an online, collections-based web-publishing platform are formidable due to the complexity and sheer mass of data and media we wish to reconcile. The paper volume comprises some 22 chapters, 327,000 words, 1161 pages, 602 figures, plus tables, references and appendices. The primary excavation database contains millions of records and hundreds of tables, in various formats – MS Access, Filemaker, spreadsheets, text files – and states of completion. The media database consists of over 15,000 images, hundreds of video clips, plus CAD drawings, illustrations, sketches and plans. Furthermore, we do not want to merely build a repository of content, but an extensible framework through which researchers, visitors and future scholars that make up the Çatalhöyük community can all make substantive contributions.

The House Lives volume is but one collective work about Çatalhöyük. There are hundreds of thousands of web pages, over 40,000 images, hundreds of videos, projects, dissertations, articles, popular books, lesson plans, and databases already out there in the World Wide Web. An overwhelming corpus of content, one of our ambitious aims for the Last House on the Hill was to find a way to link up the worlds of data within our project with the disparate universes of data out there in the Internet, or in researchers' laptops.

In order to try and make sense of all the content, a mind map (also known as 'cognitive map' [Wik09]) was used in the planning phase of the project to record our shared concepts and ideas, and to better represent the holistic and non-linear reasoning about the existing documentation and the relationships to be created between the different sources of information. The archaeological project itself was considered the central node, the original event that brought together people and things and gave origin to the complex network of information about the project. Five branches were created around the central concept in order to link up all the existing documentation, but also to open possibilities for remediating and remixing, and to allow new strategies for accessibility and usability of this rich and multiform content.
The existing databases and the online digital mirror of the book represent two of the main branches of the map. Data from the excavation, artifacts, media and the results of lab analysis are represented in numerous separate data sets that are managed by a CIDOC-CRM compatible Digital Asset Management system (see below). The branch for "remixes" represents the tools and plug-ins available on the web-platform that allow users to remix content, develop online exhibits, or build special web pages and showcases, combining digital objects in the archive with new narratives. The "media" branch represents the multiform digital content that several initiatives around the BACH project have brought online and that integrate the content of the book.

As we discuss in section 6, our greatest challenge, and an essential information management challenge for any project of this magnitude, was how to make sense of such a deluge of rich media, data and analysis. The volume is a lot of content, the primary research data and media is vastly more. Our aim is not so much to clean up the data but to link it meaningfully and make it public, to preserve through access. To do this, we needed to move out from our self-servig databases, hard drives and spreadsheets into an open data model and platform. At the same time, we have an aim and a responsibility to share...
our work as transparently as possible so that others may see for themselves what we have come up with.

4. OPEN DATA PUBLISHING

Our project has the essential aim of making primary research data safe and easy to share and to set the groundwork for remediating the past. We have come much closer to achieving these goals than in previous attempts by devising a radically simpler approach to our information management methodology, guided by a robust but easy to implement ontology. While we have made significant progress on developing a working pipeline for producing open data, the publishing platform has remained a challenge for us.

In order to promote open data, some of our key requirements include:

**Open Data / Data Portability** - The data must be in transparent and sustainable formats, easily translated into other open formats (XML, TEI).

**Open Source** - Platform should be open source and community design driven, with a solid development roadmap and sustainability plan.

**Non-proprietary, free of protection mechanisms and external dependencies** - Platform and data should be free of archaic security protocols or require other software/hardware in order to work.

**Open Rights Management** - Patents and copyright constraints can severely curtail digital content sustainability. We want anyone to be able to openly publish and remix the content of LHotH. Wherever possible, Creative Commons licensing is applied to disambiguate how content can be used.

---

For the initial platform development, we opted for the Mellon Foundation supported, open source and community developed Omeka platform [Ome09]. Omeka provides...

---

several user interface metaphors that are well suited to presenting and interacting with rich media content. A collection is a full archive of material, related by event. An exhibit is a selected set of materials from a collection, related by theme, topic or other curatorial decision. Collection administrators and collaborators can contribute, annotate, organize and manage the content using various plug-ins and tools in the Omeka back-end. End users can do almost most of these things in the front end through the MyOmeka plug-in, designed to make the process of creating an exhibit or selecting materials simple to do.

The Omeka platform has a lot of promise, but it is intended as a relatively static publication engine for collections, and like many community driven projects, will need more time to mature than we have for the project. Coming in Fall 2010, Omeka.net will offer a full cloud-hosted service that should simplify setup and maintenance for most users, akin to Wordpress.com for blogs.

Even though our approach is more aimed at the integration of text and rich media sets, it is worth to mention, as part of the international effort around effectiveness of digital publications, the Arkeotek project. This has promoted a re-writing system of scientific texts in human sciences and their edition under the SCD (Scientific Construct & Data) format. The SCD system is based on the mechanisms and foundations of the theoretical “logicist” reasoning about scientific constructs and is aimed at producing texts whose structuring helps to browse scientific constructs [Aus06] [GR04].

Since our submission to the 10th VAST International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST) in Spring 2009, we have been experimenting with the MediaHUB, an enterprise content management (ECM) platform developed by the Media Vault Program at UC Berkeley. The platform, built on the open-source ECM Alfresco repository, offers exceptional media management capabilities, including version control and extensive content model (ontology, thesaurus, logic) support. Still in early development, the platform has tremendous promise for facilitating the easy sharing and saving of research media.

Currently, we are working on the data tables and media elements using a variety of tools and applications in order to decompose and restructure the content as atomic elements that can be recomposed as required by our ontology. For example, we are using

Figure 3. Last House on the Hill in the UC Berkeley MediaHUB platform

the DropBox web service application to provide sharing and subversion control of documents and media between team members, Google docs for collaboration, Filemaker for transformation of data into OWL and XML, and Extensis Portfolio for digital asset management. While the tools we are using for convenience may be commercial or proprietary, the resulting data and media elements adhere to the highest standards for open knowledge. As we will discuss, the deluge of data are far greater challenge to digital preservation and sharing than technical platform decisions.

In the near term, we will release a web and client database application that will allow users to explore the data in its raw, atomic state. Through our partnership with the California Digital Library, we ultimately anticipate migrating all of the content - media, data, documentation, and digital volume - to their open access eScholarship portal. We discuss our data workflow in section 6 and our partnership and sustainability plan in section 9.

5. THE SEMANTIC REPRESENTATION OF DATA

Still among many archaeologists the production of systematic and exhaustive digital documentation, both visual and textual, is considered only a useful option rather than an imperative dictated by the necessity of providing explicit and objective documentation. "But in archaeology, before or alongside writing, there is the need to handle large amounts of "objective" data, and to deal with the problems of "data construction", leading to reflexive critique of even the most mundane and direct processes of observation. The methodological challenges are perhaps greater and more severe in archaeology." [Hod02]. In this statement included in the premise of "The Archaeological Process: An Introduction", Ian Hodder synthesizes a fundamental concept for modern archaeology, which is that the objectivity, multi-vocality and complexity of archaeological documentation is much better represented through the use of multimedia and hypertext.

The original aim of the LHotH project, as described above, was to harmonize and make available the final excavation report as a comprehensive digital resource enriched by related data archives and by the massive media databases produced over the years. To link media, contexts, objects, events, and people we needed to link all sources starting from the archaeological context and excavation method used. Our need was to develop an explicit path for the relationship between the find and the context, so that rather than storing data as attributes of either the object or the context, it should be possible to store data linked directly with the events to which they relate. Therefore we needed a data model that would not simply map the existing data structures, but that would be a new and enriched structure representing more conceptual and enlarged datasets. This work has been conducted largely referring to the experience done at the English Heritage's Center for Archaeology (EH-CfA) and the principles established over the time by the team that developed it [CGF*04], [MBT08]. It should be remarked, however, that we believe that some further reasoning on how the archaeological deposits and the stratigraphic relationships should be represented in the CIDOC-CRM schema is still needed.

5.1 OBJECTIVITY, RESTRICTIONS AND ABSTRACTION

Archaeologists tend to consider databases as limited structures for their "textual thinking" or their "interpretive thinking". Fields like Description, Notes, Comments, Interpretation, Discussion, abound in archaeological databases (an interesting experiment for integrating descriptive fields into formal semantic structures is in [NFH*09]). This attitude contributes in making the adoption of standard practices for documentation hard.
and slow, and dealing with mapping and ontologies is definitely not very engaging for researchers as is for computer scientists and developers. But as a direct consequence of the general diffusion of digital devices, today archaeological documentation consists of an enormously growing series of "contextual objects" that are datasets, media, reports, maps and all the information objects that archaeologists create during the research process. We are at this point in history obliged to treat this as a critical mass of information that gains its full significance if represented in all its aspects and relationships between people, actions, circumstances, time and space [ATP09].

"Historical context can be abstracted as things, people and ideas meeting in space-time" [DK06], no other expression better represents the complexity and yet simplicity of what the object of archaeological documentation is. But still a big question remains as to how to help archaeologists to represent these abstractions as formalized data. We know that to be really accessible and meaningful, the archaeological documentation needs to be normalized and requires sufficiently rich metadata. Several experiments have been conducted to model archaeological events in CIDOC-CRM based schemata [CGF*04], [DMZ06], yet we are still very far from achieving a desired integration and adopting the CIDOC-CRM into normative archaeological practice. We are still finding our way to deal with the heterogeneity of the representation, the diversity of the material and media, and with the fact that "virtually nothing can be understood or interpreted without its relation to a context of thousands of other directly or indirectly related information assets" [SAC*06].

The CIDOC-CRM ontology is one of the best tools available to date to define the underlying semantics of database schemata and document structures used in archaeology [DK06], [Cid09]. The classes and properties that CIDOC-CRM provides to explain the logic behind a database structure, to explain a database content, as well as to describe the relationships between different datasets related to the same object or event, have proved to be quite effective. And yet, the use of CIDOC-CRM in archaeological practice is still very limited. In fact, although representing archaeological information with semantic tools, that can be also machine-readable, is theoretically possible, this is still far from being a common practice. This happens for several reasons that are due not only to the difficulties of the mapping process itself or with researchers' concerns about sharing scientific data and legal practices. In the case of the CRM the required skills in managing the complexity of the system is a strong inhibiting factor. Even though, the skeptical feeling and the non-familiarity with semantic practices in archaeology is also due to a lack of publishing tools, more specifically semantic web browsers that fully support all inferences on a complex underlying ontology like the CIDOC CRM [Kum07]. Proof of this is the extreme scarcity of semantic encoded archaeological data available on the web or in any other open form. As we stressed earlier in this article, existing publishing platforms are very unfriendly for users other than computer scientists.

5.2 MANAGING AMBIGUITY IN ONTOLOGICAL MODELS

A common scenario for archaeological documentation is not a perfect world of perfectly harmonized data with homogeneous structures ready to be semantically mapped as ontological models by someone that doesn’t know the meaning of those data. Archaeological documentation is a complex universe of diverse data sets produced over the years, by different people, using also different languages, premises, and methods. Therefore the mapping process is very far from being a straightforward operation, and manually cleaning data is still a fundamental and time-consuming step before
establishing the applicability of a common underlying framework [BBC*07]. Structured archaeological information is commonly non-standardized, even within the same project. Often different categories are used to describe the same things, and shortcuts are used that are sometime very difficult to interpret. Not secondarily, archaeological databases make undifferentiated use of categories like object, object abstraction, object appellation and identifier, and object representation, which are critical in the CIDOC-CRM reasoning. In the Çatalhöyük Research Project database, expressions like "Feature", "Feature ID", "Feature Name", "Feature Code" all express the identification name or number of a feature; fields like "Name" "Initials" "Entered by" or simply "By" all give us information about the actors involved in the project, and depending on which data set contains these fields, the actors’ roles change (Figure 4). A similar situation is common also for fields related to dates and chronology of events and to location through spatial coordinates.

![Figure 4. Çatalhöyük fields, clustered by likeness](image)

The knowledge domain to be modeled was identified in the BACH excavation project and two groups of events were identified as the cores of the model: the events in archaeological time and the following formation of the archaeological record, and the recent events linked to the BACH excavation project. The ontological model started from the methodological guidelines elaborated at the beginning of the excavation of the Çatalhöyük Research Project. Each datasheet that was mapped into our ontology had its own accompanying documents like thesauri, users guidelines, and instructions of use (Figure 5). The first step was to go through all this documentation to get the exact meaning of each field in each datasheet. This process helped us in building our hierarchy of classes and subclasses, and the necessary axioms. We were very careful during the mapping process not to use shortcuts by simply extracting fields/classes from the existing database structures; instead we created more complete paths to represent the meaning,
sometimes implicit, of the data structure, not to lose any of the complexity that we wanted to represent. In the first phase of the project the model was represented in a simple spreadsheet to be easily shared within the team of archaeologists, ultimately it will be fully built in the Protégé ontology editor.

Figure 5. Çatalhöyük datasheets and accompanying instructions.

The context of the excavation has been constructed to give emphasis to the archaeological depositional context rather than to the event or action of excavating it, and it has been represented by the recursive pattern E53 Place - P 89 Falls within - E53 Place. This pattern has been used not only to represent the site sub-divisions, i.e. the spatial conceptual categories linked to the research methodology (area, space, trench, level), but also to represent the site and its depositional components of volumes and surfaces (stratum, fill, pit, floor). Although the depositional components are de facto physical things, our concern was to represent them as they are intended in the stratigraphic system and, in our opinion, they are much better represented by the CIDOC-CRM concept of "place". A principal goal of the Harris stratigraphic system is to avoid temporal assumptions during the excavation phase and "freeze" in a mathematical scheme the stratigraphic units and their relative place in the stratigraphy together with their spatial relationships [Har89], [MBT08]. In archaeology, strata are physical things that need to be described in their physical characteristics of material, color, texture, but are also places that may host other features like pits, holes, and of course findings. This is still a challenging problem using the CIDOC-CRM, as E53 Place and E18 Physical Thing don't have shared properties to represent this "duality". In fact, in the ontological model published by EH-CfA there is an added property "EH_P3: occupied" that solves the problem of linking the hosting context (E53 Place) and its materiality (E18 Physical Thing) opening the possibility of a physical description [CGF*04] (Figure 6).
As a design decision, each data source was structured according to its own unique schema and the information selected to be included in the ontology was represented as RDF triples, initially in Excel in order to facilitate the communication within the team and have feedback from the archaeologists. Mappings of table/column names and their corresponding CRM/LHotH entities were produced by an iterative collaborative process that helped in making entities and properties of the archaeological data sources clearer.

As CRM is an event based model; in order to correctly model data and connect objects and places etc., it was necessary to create intermediate 'virtual' entities (as referred to in [BMT08]) during the mapping process that did not exist in the original data sources, but were only implicit (ex. the event "excavation_of_Unit" or the entity "Unit_Place" that can be seen in the upper left rows in Figure 7). Key to RDF data entities is the adoption of a consistent convention for unique identifiers, therefore different namespaces were given to each table in the form of dot delimited prefixes (ex. in Figure 7 are represented the RDF triples for the excavation units data source, in the first column on the left in fact the prefix cus: stands for "Çatalhöyük Unit Sheet). These prefixes were adopted in temporary URI notations, allowing the reuse of the existing data source record ID values without ambiguities (ex.:http://diva.berkeley.edu/Lhoth_Ontology/cus#; http://diva.berkeley.edu/Lhoth_Ontology/css#).

In the Appendix at the end of this article there is an example of data model with declaration of main classes and properties expressed in Turtle triples. This was the initial model developed on the archaeological features data source. The latest version integrates the entities from the other data sources selected to be part of the project (Spaces, Units, Finds, etc.). We are in the process of selecting and testing data extraction utilities, the actual publication of the data though still remains a big challenge given the quantity and variety of media associated with these data. The experience with LHotH confirmed that a
semantic mapping very well meets the need to integrate all data produced by the various disciplines that contribute to an archaeological project.

The CIDOC-CRM was an obvious choice, be it an ISO standard for event based modeling of Cultural Heritage information. The high level conceptual categories of the CIDOC-CRM proved, in fact, to be very relevant to the complexity of this archaeological project and its related activities, although more remains to be done to make it easily applicable during the research process, and, as stressed earlier, still doesn’t exist a publishing platform that can display a complex and massive content through a friendly interface.

6. IMPLEMENTATION AND ARCHITECTURE

The immediate aim of the LHotH architecture is to implement a lightweight but effective semantic mapping in order to make it possible to define portable data streams that can be easily moved, transformed and repurposed without losing their underlying meaning. This is the goal of the data portability movement, which has arisen as the walled gardens of social networking platforms such as Facebook and Twitter have demonstrated new challenges in this domain [Biz09]. Data portability assures that preservation and privacy concerns are met while providing remarkable flexibility to how data streams can be managed and structured. In order to achieve this, we needed first to atomize all of the data sources, guided by the ontology described in section 5.

Figure 7. Class and Property assertions in the LHotH ontology.

The CIDOC-CRM was an obvious choice, be it an ISO standard for event based modeling of Cultural Heritage information. The high level conceptual categories of the CIDOC-CRM proved, in fact, to be very relevant to the complexity of this archaeological project and its related activities, although more remains to be done to make it easily applicable during the research process, and, as stressed earlier, still doesn’t exist a publishing platform that can display a complex and massive content through a friendly interface.

6. IMPLEMENTATION AND ARCHITECTURE

The immediate aim of the LHotH architecture is to implement a lightweight but effective semantic mapping in order to make it possible to define portable data streams that can be easily moved, transformed and repurposed without losing their underlying meaning. This is the goal of the data portability movement, which has arisen as the walled gardens of social networking platforms such as Facebook and Twitter have demonstrated new challenges in this domain [Biz09]. Data portability assures that preservation and privacy concerns are met while providing remarkable flexibility to how data streams can be managed and structured. In order to achieve this, we needed first to atomize all of the data sources, guided by the ontology described in section 5.
Figure 8. Implementation ontology: Place, people, things and media bound geo-temporally.

Our implementation ontology elevates all media and data elements to the same level of relevance (class), and makes it technically painless to curate the entire universe of rich media and data despite its enterprise level complexity and sheer mass. The system is based on the CIDOC-CRM ontological model and aimed at describing the complex network of entities, relationships and properties that together constitute the archaeological knowledge. The premise of our database structure is a framework of things, people, places, and media, meeting in space and time. Archaeological documentation [MEDIA] are the interpretations of material culture in relation to the historical events that produced them, and can be described in terms of events and processes. We have focused on defining the essential entities through unique properties, the relationships of links between any two entities, and continued this process until we defined all of the potential events that took place in the excavation, as well as after (post-excavation and publication).

The first order product of our open data workflow has been achieved using an application we developed, lovingly described as the meat grinder. The LHotH data universe is an RDF triplestore of unique entities, relationships and properties, distilled from the data sources, along with a managed set of archival media – images, documents, video, drawings, and databases. From these elements we can produce any variety of outputs – XML, ATOM, TEI, OWL – and implement in a relational database, preferably in a graph database or native XML. We are currently exploring the SESAME open source repository.

We had several principles for the data-structuring phase of the project:

Principle 1: Interoperability – The Çatalhöyük Research Project continues and we wanted to make it as easy as possible to update the LHotH atomic database and also return to the research project verified data and media in the formats required by the CRP.
Principle 2: **Disambiguation** – We define our pipeline of data processing as origin | transformation | destination. It is critical to be aware of how the data will be changed as it moves through the pipeline. As described in section 5 and figure 4, entities can be described by quite similar properties (fields) and the same property may have quite different meaning, depending on its context and use (data source). Are we talking about the same thing, and in the same way? Once we identified all of the entities – people, places, things and media items – we went through the painstaking process of assuring their uniqueness. The result is a unique entities table, validated by the project director.

Principle 3: **Simplicity through transparency** – Our super-class model [PEOPLE] [PLACES] [THINGS] [MEDIA] [SPACE-TIME] radically simplifies the implementation of the CRM for non-information architects. As shown below, this model is not unique to archaeology and can be applied to virtually any collection or project. We have attempted to put as much of the mapping drudgery ‘under the hood’ during daily practice, but we maintain complete transparency of each mapping step within the system. In fact, an early design decision was to embed full provenance and documentation inside the database, text items and media files, even though most end users will never see it. For preservation and access, it is essential.

6.1 DEFINING PEOPLE, PLACES, THINGS AND MEDIA THROUGH EVENTS

As Doerr et al. argue, the CIDOC-CRM provides a nearly generic information model for handling cultural heritage events, documents, places and people [DI08]. While the model provides the classes, instructions and ligatures needed to model virtually any data, specific case studies and CRM ‘wiring diagrams’ are needed to effectively implement the model. We have taken up the challenge of designing a radically simpler approach to the implementation of the CRM classes that does not dumb down the complexity of the data, but instead focuses our attention by constraining the types of entities through the use of five super-classes.

Figure 8 depicts sample entities from Çatalhöyük. Our aim is to define all of the unique entities, properties and relations involved in the BACH excavation and subsequent documentation activities:

[PEOPLE]: Data models often overlook the involvement of people as actors with varying levels of involvement in the events that create the data. To the highest degree possible, we have associated people with the events, describing their roles during the excavation, initials and pseudonyms, their positions now, interests and years on the project. There are 74 people who worked on the BACH excavation and many more who have been involved in the volume production, related multimedia and publication projects. The list of people could be extended infinitely as future researchers explore the database, make observations, or remix the media in new works.

[THINGS]: For the purposes of this project, we have defined things as the archaeological entities; physical stuff such as finds, skeletons, depositional units. While a film slide is a physical thing that the CRM is perfectly equipped to define, where appropriate we use CRM classes for physical media objects in our [MEDIA] super-class. To keep things simple and in scope, archaeological stuff are the things we looked at, a relatively short list of several thousand.

[PLACES]: Spatial contexts where events take place. We adhere to the CRM definition of place [E53] but add the relations defined in the Harris Matrix for the single
context recording system deployed at Çatalhöyük (see discussion is section 5), mapped to
the CRM. Places are the spine of our system. There are 1356 places uniquely defined for
the BACH project, hierarchically defined as


Each place has been individually validated by the project director (Ruth Tringham)
and cross-correlated with other names and uses (eg. DigHouse/Dig House, Experimental
House/Replica House). One remarkable challenge has been relating place records in the
database or asset management system to the volume text. Take this example from chapter
4 of House Lives:

Building 3 was excavated as Spaces 86, 158, and 201. At the time of its construction,
the entire building was used as one open space (S.201).

While there is only one Space 86 in the BACH area, there are an additional 34 Units
that contain 86* - 8635, 8636, 8637, etc. There is also a Feature 158 and chapter 4
contains a Figure 4.158. We built an algorithm for grinding through the volume texts, and
then linked up the places, things and media to the texts, paragraph by paragraph, using
standard CRM relationships.

An outstanding semantic challenge awaits us for implicit references to places and
things. For example, specialist report may not at all refer to an archaeological feature
directly:

…not very compacted lenses of oven rake-out and burnt and non-burnt aggregates
accumulated in thinner layers on the roof in W255, less than 1-3cm thick (Figure 7.1).

W255 refers to a sample taken of Feature 157, the collapsed roof depicted in photo
PCD0013_088.tif in 1998. The photo log sheet does not mention the sample, but because
W255 is in Feature 157, once we have made the link, we can connect any mention of
W255 to entries about Feature 157, including associated photos, drawings and diary
entries.

[MEDIA]: Our general category for all of the project documentation. In strict CRM
terms, a portion of these media are physical things such as hard disks, videotapes,
drawings, CD-ROMS, but the majority of the content is digital. The richness of the media
record for LHotH is as extraordinary as it is overwhelming. There are some 44,000
images, video clips, illustrations, sketches, plans, 3D models and texts relating to the
BACH project. In recent years, the CIDOC working group has made great progress on
considering digital media and provenance for the purposes of long-term preservation
[Doe08]. We have applied our meat grinder to our media items and have been very
successful at linking them to the people that took them or are in them, the places where
they were taken, the things depicted, with the principles of empirical provenance in mind.

As with the Places problem, file renaming is a common occurrence. Fortunately, the
BACH project has kept a remarkably accurate audit trail of filename transformations
throughout its life. Small links can have huge gains. For example, Figure 7.1 mentioned
above relates to PCD0013_088.jpg, a derivative of PCD0013_088.tif, which is the digital
surrogate of a Kodak Ektachrome slide taken by Michael Ashley and digitized in 1998.
There are 133 digital originals that contain Feature 157. Now that the links are in place,
we can derive a much richer story of the feature, including its archaeological phase,
mentions in other chapters, related units, people who worked on it, all from just the photo
captions alone.

[SPACE-TIME] Finally, we take full advantage of the power of our ontology for
defining events in space and time. Any entity can have where, when, who, what, how, and
why information associated with it, either directly or indirectly [Add08]. By linking up
the entities, properties and relations through events, we can associate the geo-coordinates of a feature, for example, with a media item depicting it. Feature 157 has a centroid of Grid X 1055.2 and Grid Y 1186.05. PCD0013_088.tif is a photo of Feature 157 and can inherit its location. We can convert these coordinates to UTM or WGS84 and publish a photo of Feature 157 as .kml in Google Earth, even though the original photo was taken in 1998 and no location data was recorded.

We have found this cross-linking of data to be critically useful in the writing and fact checking of the House Lives volume. For a project that spans seven years and is being published more than a decade from its initiation, leveraging multiple lines of evidence not only helps us to produce a better, more accurate accounting of the excavation, but is also the central aim of semantic integration.

7. UNBURYING THE PAST SEMANTICALLY: REVEALING BURIAL FEATURE 634

As a test case by which to measure the effectiveness of the Last House on the Hill in user engagement and research potential over the paper format, we have chosen to focus on a single feature: Feature 634 – the burial of a mature woman under the north-central platform (F.162) of Building 3 in the BACH Area at Çatalhöyük. This same platform in fact was cut by three additional burials whose pits, together with that of Feature 634, broke almost the entire surface of this low max 25 cm high 2x2 m. plastered platform. Such group burials under occupation floors were known from many other Neolithic buildings at the site, but this was the only one in Building 3. Feature 634 was first recognized as a probable burial pit on July 29, 2000, and was then excavated for at least 24 days until its final destruction in June 19, 2001. During these 24 days it was the
subject of intensive retrieval of artifacts and samples as well as documentation in various formats.

Feature 634 is mentioned in almost every chapter in the House Lives paper monograph, especially the excavation report (Chapters 2 and 4) and the Human Remains report (Chapter 11a), and is documented in 25 images (few of which pertain only to F.634). It also figures indirectly in other chapters because of its contained basket phytoliths (Chapter 10), flaked stone (Chapter 15) and beads (Chapter 17) and the practice and personnel involved in excavating the burials (Chapter 3). It is illustrated in 23 of the book's figures. It is recorded exclusively in 10 units but is significant for 47 others. Each unit has large amounts of quantitative and qualitative data recorded on its unit sheet, along with associated data from flotation samples, micromorphology and other soil samples, as well as artifactual finds and detailed skeletal sheets. The units are further documented by - in the case of Feature 634 - 600 high resolution color images, over 40 video sequences, QTVR movies, 11 detailed field drawings, and 40 detailed diary entries by the field directors. And we should not forget the 16 people engaged in various roles in the excavation of Feature 634. When the interpretive and descriptive texts of House Lives are integrated online with this full inventory of original data and documentation on the people, places, things, and events related to this feature, the increase in the potential (but also the challenges) for the public (whether professional or laymen) to understand, evaluate, and interpret the investigation of this feature is enormous. This potential grows enormously not only when we realize that these data and media are just one part of similar datasets for all the other spatial elements of the BACH research project, but also that these data and media are integrated and made accessible with the architecture.
described above, and with standards and protocols that enable its utilization for the long-term.

Figure 11. Feature 634 in context with related features and units.

8. THE CONCEPT OF "DATABASE NARRATIVES"

We have shown in this paper that the mass of archaeological documentation gains its full significance for a study of the past if it is represented as the relationships between people, actions, tasks, and the contingencies of time and space, all of which contribute to the creation of the archaeological record. Narratives that represent these relationships can be drawn out of the database through the filter of the alternating perspectives or standpoints of people, places, things, and media, thus enabling the recontextualization and remixing of the content that resides the database. Thus the ultimate aim of the Last House on the Hill is to have both archaeologists and a broader public use our data and media in creative and productive ways to think about both the past and the present. In this respect we have been inspired by the concept of both database narratives [Man01] and recombinant or digital history [And].

Steve Anderson recognizes two directions in which historiography has embraced digital technology. On the one hand is the idea of amassing the “total” historical record of events through accessible networked interoperable databases, creating history that is as definitive as possible. On the other hand “digital technologies have enabled strategies of randomization and recombination in historical construction resulting in a profusion of increasingly volatile counter-narratives...and histories with multiple or uncertain endings”
Both these practices are legitimate pathways to creating narratives about the past. The idea of recombinant history, however, has been a guiding principle of the Last House on the Hill project, reflecting the fragmentary open-ended nature of memory and history drawn from a database with a structured architecture.

At the core of our project structure, groups of events can be identified (events of a different scale from the events that bind the database itself): these are events in archaeological (Neolithic) time and the subsequent formation of the archaeological record, and the recent events linked to the BACH excavation project. Out of our databases an endless and fragmentary configuration of people, places and things can be created as vignettes or narratives, seen as juxtapositions of image and text or video montage or text alone - narratives of people, places and things. Some "narratives" will be sparked by observing patterns in the spatial or chronological occurrence of artifacts in the excavation record, others by juxtaposing conversations on a video with narratives from a text, or by images with diary entries. Some of these narratives will be descriptive or informative, following Anderson's (in press) first historiographic option - what Bolter and Grusin refer to as 'respectful remediation' [BG99], in which texts such as those in House Lives are enriched by use of digital photography or video in a pleasing design, but with little other creativity. Such remediation of the BACH Area (including Feature 634) can be seen by the examples of Remixing Çatalhöyük, or a presentation to the European Archaeological Association in 2002 on using the Harris Matrix in platform Feature 162. In the Last House on the Hill, the architecture of the data and media also encourage a more 'radical remediation' [BG99] which encourages a user to critique and improve on other media (and narratives) in the process of mediation. In this case, an explorer of our data re-uses and re-contextualizes our excavation products to create in their own narrative an alternative interpretation or 'reading' of the data, challenging the apparent authority of the 'expert', even using such tools as irony and satire [JT07], [TAM]. We have already created some 'remixes' of this kind that are based on Feature 634: RAVE: Requiem for a 'boneyard'; Dido's Lament; Remixing Dido's Lament; the F.617 movie; and the time-lapse sequence of the excavation of F.617 that was created by the Science Museum of Minnesota as part of a multi-year National Science Foundation grant. A third possibility would be to enable software to create the remixes according to parameters set by the user, along the lines of Lev Manovich's Soft Cinema in which "the software edits movies in real time by choosing the elements from the database using the systems of rules defined by the authors" [Man05]. This would create many different narratives from the same data.
We have shown in our test case – the burial Feature 634 – that one relatively small feature can generate an enormous amount of data and documentation in a variety of formats. There are limitless strands that can be followed to explore these data about of people, places, things, and events that are entangled in Feature 634, and to use them creatively to produce new narratives or readings: the baskets, the bodies, the stratigraphy representing life-history, the practices that surround treatment of the dead, the people who specialize in their excavation, the procedures of their excavation, the story of Dido with her broken ribs and dislocated hip, and so on. In Chapter 22 of House Lives, entitled "the Senses of Place at Çatalhöyük", we could do little more than suggest such narratives as these that might go beyond the strict empirical boundaries of the excavated data as tantalizing but disembodied fragments. In the Last House on the Hill, we can use the full power of digital technology to take these fragments - vignettes - of created knowledge and combine and recombine them into an open-ended but always accumulating history of the BACH project, that is itself a fragment (often forgotten) of the larger Çatalhöyük Research Project, that leads onto the history of Anatolia.

Extracting one such narrative starts with a person in the project time - Lori Hager excavating the skull of Feature 634 and being filmed by Ruth Tringham. The narrative draws on data on Lori’s Biography, video clips of the excavation, images of excavating, drawing, units, images, her human remains colleague, Başak, everything captured in Ruth’s diary of the event as well as her video: Friday, August 18 2000, and then remediated by Lori’s words in chapter 11 of House Lives and a video interview in which Lori remembers what she sensed when excavating this skull. This same event is
transported through the medium of an interactive Harris matrix representing the sequence of events surrounding Feature 634 to the prehistoric event and a prehistoric actor, the story of the woman we call Dido.

Figure 13. Some of the countless narratives of events that can be channeled from the database of Feature 634 in Building 3 through the filters of people, places, things and media in the Last House on the Hill.

9. COLLABORATIONS AND CONCLUSIONS

On May 28, 2009, a small software development team in Sydney, Australia, the same development team that gave the world Google Maps, announced Google Wave, a singular communication platform that will transform the way humans interact with each other digitally [Ras09]. In the next few years, recombinant history will be a way of life in the cloud, as the lines drawn between e-mail, text messaging, blogging and media management are washed away and we simply communicate with each other, instantly, globally. Waves will thrive on high value information, database narratives, and whole cloth event streams than can be dragged and dropped into conversations as easily as photos are now handled in iPhoto or Picasa. In June, UC Berkeley researchers announced the development of a new storage nanotechnology that could preserve information for billions of years, potentially eliminating the digital preservation crisis we are still facing [Hea09], [Ash08]. In the wake of such innovations, it is high time to be casting away the limitations that are self-imposed by traditional publishing. We really can save and share it all. There will be challenges and setbacks ahead, but we are can see no other way than forward.

Part of what we, as archaeologists, have been striving for is that the results of our archaeological investigations do not themselves become part of the archaeological record,
languishing and rotting in a forgotten place. We believe that the digital formats have made (and hopefully will continue to make) a significant contribution to the long-term preservation of archaeological archives and their accessibility as sources for information and recontextualization into new knowledge. Such contributions are not inherent in digital technology but have to be enabled by the archaeologists who are providing the content and the digital architects who help them make it happen.

![Figure 14. LHotH data visualization application.](image)

In conclusion, we have high aims for the project, and while there is a lot of work to go, we would like to mention a few important accomplishments:

**Our primary research data is safe and easy to share** – All of the project content, data and media, are digitally curated, validated and redundantly backed up. All content has clear licensing guidelines (Creative Commons, wherever possible) and are in formats that are easy to share and reuse. This achievement was particularly demanding as it required going over tens of thousands of data using a variety of tools and applications in order to decompose and restructure the content as atomic elements that could be recombined as required by our ontology. A meticulous process of data review was required to identify and cleanse incomplete, incorrect, or inaccurate records, which is a very natural consequence for a field project spanned over many years and performed by many different actors even if based on a shared methodology. We are in discussions with the California Digital Library’s Digital Preservation Repository for the long-term curation of this immense body of work.

**Easy-to-use event centered implementation framework** – We have used the framework described in section 6 on several projects with compelling results. Our next
step is to complete and freely publish the documentation so that other researchers can see if this methodology will benefit their projects.

Over 40,000 media items, spanning a period of time from 1998-2003, are now linked to the excavation events and people that participated in them. Instead of one paper volume with black and white images, we are able to share a far richer accounting of the excavation, including all of the original images, video, sketches and ancillary data.

The Last House on the Hill project is made possible due to the democratic decision made by the Çatalhöyük Research Project to release all data and media under Creative Commons open licensing, freeing the content to be shared widely with little intellectual property restrictions. We are grateful to the team for this choice and for all of their contributions. More details can be found on the project site, www.catalhoyuk.com.

The Cotsen Institute of Archaeology has agreed to partner with the LHotH team to take the deep dive into this new paradigm of holistic digital publishing, through collaborations with the California Digital Library e-Scholarship program and the UC Berkeley Media Vault Program. We are using this occasion to launch ArchaeoVault.org, a digital preservation and access resource for archaeological projects to be perpetually stored and shared.

The Last House on the Hill can be found at www.archaeovault.org/lhoth.

REFERENCES

9: 25 ● M. Ashley, R. Tringham and C. Perlingieri


[OKF08] http://www.openendination.org: Open Knowledge Definition 1.0


[Tri04] Tringham, R.: Interweaving Digital Narratives with Dynamic Archaeological Databases for the Public Presentation of Cultural Heritage. In Enter the Past: The E-way into the four dimensions of Cultural heritage - CAA2003, edited by W. Börner and W. Stadtarcheologie, pp. 196-200 (full version on accompanying...
APPENDIX 1

LHotH main classes and properties declaration schema expressed in Turtle triples.

# BACH - Çatalhöyük Ontology by Cinzia Perlingieri - UC Berkeley (CA) 10.26.2009
@prefix crm: <http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdf#>.
@prefix lhoth: <http://diva.berkeley.edu/LHotH_Ontology.rdfs#>.
@prefix rdf: <http://www.w3.org/1999/02/rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

lhoth:LHotH4.Team_Members_Role rdf:subClassOf crm:E55.Type.
lhoth:LHotH6.Catalhoyuk rdf:subClassOf crm:E44.Place_Appellation.
lhoth:LHotH11.Area_Name rdf:subClassOf crm:E44.Place_Appellation.
lhoth:LHotH15.Space_Name rdf:subClassOf crm:E44.Place_Appellation.
lhoth:LHotH19.Unit_Name rdf:subClassOf crm:E44.Place_Appellation.
lhoth:LHotH34.Wall_Sheet rdfs:subClassOf crm:E73.Information_Object.
lhoth:LHotH43.Sheet_Date rdfs:subClassOf crm:E50.Date.
lhoth:LHotH44.Archaeologist rdfs:subClassOf crm:E55.Type.
lhoth:LHotH47.Media_Date rdfs:subClassOf crm:E50.Date.
lhoth:LHotH52.Team_Members crm:P14-1.in_the_role_of lhoth:LHotH3.Team_Members.
lhoth:LHotH53.Team_Members crm:P14-1.in_the_role_of lhoth:LHotH4.Team_Members_Role.
lhoth:LHotH64.Feature crm:P87.is_identified_by lhoth:LHotH15.Space_Name.
M. Ashley, R. Tringham and C. Perlingieri


Received January 2010; accepted XXXX