Extending the ETL Process Through a Working Data Interchange Graph and Understanding Different Realities of Data

Brian Ballsun-Stanton, Shawn Ross, Adela Sobotkova, Penny Crook

UNSW, brian@fedarch.org
UNSW, shawn@fedarch.org
UNSW, adela@fedarch.org
LaTrobe, penny@fedarch.org

ABSTRACT

Data is useful in proportion to its availability. A significant problem in research is data housed in non-compatible silos, which restrict access through storage format rather than intention. The Federated Archaeological Information Management Systems project aims to address this problem. Besides building applications for digital data collection and analysis, this project will develop interoperability approaches connecting generated datasets with key existing online archaeological data repositories through a universal web API.

Most existing archaeological field databases have few points of commonality. There is no consistent universe of discourse, much less common ontologies or architectures. These databases have been designed to meet the unique requirements faced by individual projects; not to communicate with each other effectively. A few online tools and repositories (e.g., Heurist, OpenContext, tDAR) have made provisions for “loose coupling”, where the results of a query can be exported as an XML payload via RSS or similar means. Interoperability is, however, limited in the current universe of archaeological data.

People have different realities of data. These realities of data shape their requirements and programming decisions when implementing software. These realities also shape how users perceive the data being presented by interfaces. This cycle reinforces and alters meaning, creating local silos of what data is “supposed” to be. Each of these database management systems renders the subjective reality of the programmers, recording only that data deemed important and necessary from the perspective of the requirements used to generate them. Our project will create not only Extract, Transform, and Load [ETL] processes to transform the data generated by these diverse systems, but work towards a common data interchange graph such that new systems may be made communicable with these through a simple protocol. Because of our inclusion of philosophy into the ETL component, we believe that we can achieve more than a simple “loose coupling” as OpenContext and tDAR are trying for.

FEDERATED ARCHEOLOGY AND THE DATABASE

The Federated Archaeology project aims to create a means by which archeogolocial data consumers and data producers may “federate”. Federation is the protocol, means, and intention of different technological systems to communicate with each other via a shared standard. In this case, we will unite data-producing field units and their transactional databases with the common warehousing options that exist.

By generalizing this process beyond a simple ETL\(^1\), we can build a staging database and necessary software protocols so that new entrants into the archaeology field may participate in this federation without additional effort from us. We currently are exploring federation options to integrate tDAR, Open Context, Heurist, New South Wales Archaeology Online, state cultural heritage registries, and so forth. A properly constructed protocol should be able to translate between these databases and their realities of data. By encoding common validation problems: both technological, typo, and meaning shift, into standardized rules; it will reduce necessary human involvement in automated processes as well as increase the fidelity of the final product. The novel incorporation of inspecting and translating meaning-shift between databases should provide significant utility.

Federation between databases is a difficult project at the best of times. There exist no standards to describe the nature of data within a database. While data types and lookup tables certainly constrain the domains and some values of tables respectively, the constraints do not apply beyond a technical level. There are significant epistemological questions (what does this table care about?) and ontological questions (what is the reliability of these observations? Do they represent a reality or are they fictions for bookkeeping purposes?) that are assumed and implemented without a second thought by designers.

Every database must be limited in scope, as no database can create a 1:1 map to reality. The chosen and intentional limitations to scope, either at first conception or as altered through later revisions is called the “Universe of Discourse.” Seldom, however, is the universe of discourse documented. In the act of

\(^1\) The ETL process as described by Kimball and Inmon is not exactly “simple.” However, the process is a fundamentally technological exploration that does not look at semantic or philosophical loading of the words. Incorporating philosophical elements into the process can only help to enhance meaning translation, instead of simple data translation.
documenting assumptions about the limits and nature of the database, the first steps to Federation will be achieved.

Databases can only serve as stores of intentionally created data. It is impossible to accidentally archive data inside a database. Therefore, the act of recording not the thing, but the symbols representing the thing, is what creates the inherent subjectivity of a database. This subjectivity is present in the encoding regardless of the subjectivity or objectivity of the data itself. Another crucial step necessary for federation is to look at the practices of recording within the databases we hope to federate so that we may learn the nature of the data through example. By looking at how the databases are used and what meaning the relationships within the data attach to the contextualization of data, we may then document that and apply necessary transformations to both the data and its relationships.

REALITIES OF DATA

The essence of our insight necessary to produce a true, working, federation is rooted in the understanding that people have different philosophies of data [1]. Given that these philosophies provide an epistemological and ontological framework for people to understand the nature of data, different philosophies therefore result in both different database designs and different interpretations of data stored within that database.

Different realities, or philosophies, of data impact the design of the database through choices of meta-data storage, the universe of discourse, and normalization. If data is considered to have a “reliability” indicating how valid or true it is, that reliability may be stored in one database but not others. Data may be said to stand alone, or require context to turn it into information. The very name of the “fact table” in a data warehouse implies a huge host of assumptions both of the nature of data (can it contain facts) and questions of the philosophy of science exploring what exactly a fact is. Furthermore, as data warehouses are designed from transactional databases, the warehouse designers are not only applying their own reality of data to whatever set of data they are manipulating, but they are mutating the design and the intent of the original, transactional, programming team.

We have identified three realities of data operational in the recording of archaeological (and other) information, which we must account for during data interchange graph design. Subjective recorded observations posit that data is a function of humans and human-designed sensors. It is an encoded symbol representing a human interpretation of event². Therefore, data must be encoded with an appropriate context so that the significance and nature of that human observation may be accurately transmitted. Measured objective values posit a true underlying reality. In contrast to the other philosophies, data exists in the relationships of atoms to each other and is not created by humans. Data is the measurement of the real and should be self evident up to the limit of the tools used to measure. In this sense, what must be encoded along with the data are the frailties of the measurement technique. Data of this nature encoded into a database will have far different significance to the designers than the prior reality of data, thereby changing the design and impacting the ETL process. Encoded human communications are different in nature and kind: data is that which is stored within a computer. Databases with this belief will provide for access and retrieval of the data, but little “transformative” ability, as there is nothing “there” to be transformed. Creating our federation protocols with these different philosophies of data in mind will help us to avoid the ETL process changing the meaning of the translated data.

A DATA INTERCHANGE GRAPH

With the above databases in mind, and with questions of the philosophy of data, we posit the need for a data interchange graph. This graph will provide the basis for the federated archaeology tool as it provides a way of translating data to and from a common format. This format extends beyond a simple technical specification and into a way of exchanging meaning.

The data interchange graph will look much like a data flow diagram. The critical difference, however, is that each vertex which “transforms” the data will provide a list of prompts to the designer. These prompts should provide ways to encode the context and metadata present in both the documenter’s mind and discoverable within the database itself.

By performing this methodology and using a process to iterate through their tacit understanding of the model and turn it into a fully explicit schema, the graph will then provide a a means to create the necessary regular expressions and SQL for data transformation. Beyond that, a system for translating data between the fundamental database management systems that these projects are implemented upon will also provide a great boon to the academic community: providing more ways to compare and analyze now commensurate data against each other.

REFERENCES


² People with this philosophy of data tend to use data as a singular and “data point” instead of datum.
ABOUT THE AUTHORS

Brian Ballsun-Stanton

Dr. Brian Ballsun-Stanton is a Philosopher, Information Technologist, and Game Theorist exploring the Philosophy of Data. He has a Ph.D in Philosophy from The University of New South Wales in addition to a MS & BS in Information Technology from the Rochester Institute of Technology.

He has developed a new methodology (The Social Data Flow Network) to explore how individuals in the field understand the nature of data. Currently, his research is exploring the social construction of technology, focusing on the user-driven change of technological tools, in addition to his research of data warehouse ETL strategies for disparate data sets.

Beyond philosophy, Dr. Ballsun-Stanton has experience in haptic research, specifically the use of gesture control of human computer interfaces and an abiding interest in robotics as seen in his flying robotic cable-array manta-ray.

His interests include the academic study of Role-Playing games, the exploration of how science fiction literature transforms reality, and the social consequences of a technological world outpacing society’s cultural assimilation of its consequences.

Shawn Ross

Dr Shawn Ross (Ph.D. University of Washington, 2001) is currently a Senior Lecturer in Ancient Mediterranean and World History in the School of Humanities and Convener of the Archaeology Minor at the University of New South Wales, Sydney, Australia. Dr Ross’s research interests include the history and archaeology of pre-Classical Greece, oral tradition as history (especially Homer and Hesiod), Greece in its wider Mediterranean and Balkan context, and the application of information technology to the humanities. Recently, the focus of Dr Ross’s work has shifted to fundamental archaeological research in Bulgaria, applied to the history of Greeks, Macedonians, and Thracians in what is today northern Greece and Bulgaria. Dr Ross is a Research Associate at the American Research Center in Sofia, Bulgaria, and currently supervises the Tundzha Regional Archaeological Project (http://www.tundzha.org), a large-scale archaeological survey and palaeoenvironmental study in central and southeast Bulgaria. During 2012-2013, Dr Ross will direct the development of information management systems for archaeology as part of a large National eResearch Collaboration Tools and Resources infrastructure grant.

Adela Sobotkova

Dr Adela Sobotkova completed her PhD in Archaeology at the University of Michigan (Interdisciplinary Program in Classical Art and Archaeology) in January 2012. Adela used the surface survey data from Tundzha Regional Archaeological Project (TRAP), Bulgaria, as the basis for her dissertation on the evolution of settlement patterns and polity in Thrace. Her interests include the history and archaeology of the Black Sea region, theories of state formation, integration of different categories of data and the application of GIS and remote sensing to archaeology. Adela is one of the supervisors of TRAP; she manages the day-to-day operations of fieldwork, helps to develop project methodology, administers the GIS and remote sensing databases, and leads field teams. As of June 2012, Adela has joined the School of Humanities at UNSW as one of the coordinators of the National eResearch Collaboration Tools and Resources infrastructure grant, while also working on the final report on TRAP 2009-2011.

Penny Crook

Penny Crook is a historical archaeologist, consultant and honorary research associate with the Archaeology Program at La Trobe University. She specialises in urban assemblage analysis and material-culture studies. She designed two customised relational databases to store archaeological and historical data derived from the ARC Linkage-funded ‘Exploring the Archaeology of the Modern City’ project. She has collaborated with Victorian eResearch Strategic Initiative (VeRSI) and the La Trobe eResearch Office on the Australian Historical Archaeological Database (AHAD), funded by the Australian National Data Service (ANDS). She is currently a Research Fellow at La Trobe, collaborating on the NeCTAR-funded Federation of Archaeological Information Management Systems (FAIMS) project led by the University of New South Wales in collaboration with Intersect, VeRSI and archaeological departments and private practices across Australia. She has published several papers and reports on urban archaeology, consumption studies, artefact cataloguing and database design.