Type: Article

Title: **Creating eResearch Tools for Archaeologists: The Federated Archaeological Information Management Systems project**

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

Institution: University of New South Wales, La Trobe University

Email: [shawn@fedarch.org](mailto:shawn@fedarch.org), [adela@fedarch.org](mailto:adela@fedarch.org), [brian@fedarch.org](mailto:brian@fedarch.org),[penny@fedarch.org](mailto:penny@fedarch.org)

Contact Telephone: 04 01 758 300 (Shawn Ross)

# Abstract:

In this article the Federated Archaeological Information Management Systems (FAIMS) project presents its stocktaking activities and development plans for the the creation of new digital infrastructure for archaeologists. The project, a National eResearch Collaboration Tools and Resources (NeCTAR) funded initiative, aims to develop tools facilitating the creation, sharing, re-use and dissemination of high-quality digital datasets for research and cultural heritage management. FAIMS has engaged in extensive stocktaking activities with archaeologists and related professionals, the results of which have shaped the development plans. Project development is focusing on highly customisable mobile applications for data collection, web application for data processing, and an online repository for archiving and disseminating data, with provisions for creating semantically and well as technically compatible datasets embedded throughout. Data exchange using standard formats and approaches ensures that components work well together, and that new, externally developed tools can be added later. Our goal is to create a digital product that improves the availability of compatible archaeological data, but remains something that archaeologists will actually want to use.

# A Review of Archaeological Data Management

Archaeology stands on the brink of major change, propelled by converging technologies. In recent years, a revolution in mobile devices has occurred, epitomised by Apple’s iPhone and iPad, but now extending to a vast array of devices built around Google’s open-source Android operating system. At the same time, robust open-source ecosystems for data management, CAD, mapping and GIS, and other features of interest to archaeologists have matured. Within the discipline, high-quality online archives and publication services for archaeological data have been implemented. Taken together, these advances provide fertile ground for the development of archaeological information management systems that shepherd data from digital creation, through editing and analysis, to online archiving and publication of reusable datasets. Such an ecosystem would improve the scope and rigor of archaeological research by facilitating reinterpretation, promoting regional and comparative studies, and broadly contributing towards the repurposing of data.

Despite the obvious potential, however, archaeological information management has not reached its full potential. Archaeology suffers from “small science” data problems (e.g., diverse and idiosyncratic datasets, customised methodologies and recording systems, lack of core data standards, limited budgets, etc.), that inhibit the production and dissemination of high-quality, compatible data. \cite{Kansa2010f, Kansa2010c} In 2006 a US National Science Foundation-funded workshop concluded “for archaeology to achieve its potential to advance long-term, scientific understanding of human history, there is a pressing need for an archaeological information infrastructure that will allow us to archive, access, integrate, and mine disparate data sets” \cite{Kintigh2006a}. These conclusions were echoed in the “Policy Forum” of the journal Science \cite{Snow2006b}, which observed that “archaeological research remains a mosaic of parochial efforts…[r]esearch on large geographical areas is particularly difficult at present”. The authors suggest a way forward built around data standards and translation protocols, concluding that “Federated databases and ontology-based database integration…provide the means for coordinated data management”. The population of online archives has also been disappointing, in part because submission of data to them has not been integrated into the data lifecycle \cite{Fleischer2011, Optimized2011}. Progress has been made, largely in the realm of discrete tools or resources, but there has been little attention paid to the development of a rational data lifecycle, and less of a convergence towards interoperability and shared standards has occurred than might be hoped \cite{Kansa2011a, Ross2009a}.

This paper presents the activities of the National eResearch Collaboration Tools and Resources (NeCTAR)-funded Federated Archaeological Information Management Systems (FAIMS) project. This project examines the entire archaeological data lifecycle for opportunities to improve efficiency and enhance the quality of data, with the ultimate aim of producing and disseminating compatible datasets - but doing so in a way that accommodates current archaeological practice and is broadly acceptable to the archaeological community. To accomplish these goals, we present a suite of modular, federated tools, focusing on digital data creation, online archiving, and data portability.

## Project Background

The FAIMS project was initiated to develop robust information management systems beyond ~~what~~ the scope of individual research projects or organisations. What we had initially planned as a modest project to produce custom mobile device applications and databases quickly grew, leading to an (unsuccessful) ARC Linkage Infrastructure and Equipment Fund (LIEF) application in 2012. The application was refined, and the network of participants expanded over the course of two additional (unsuccessful) grant applications, eventually leading to an ambitious proposal emphasising the development of open source tools to manage archaeological data from digital creation through analysis to online archiving and publication. This revised and expanded proposal formed the basis of a successful NeCTAR eResearch Tools application underpinning the development described in this article.

NeCTAR awarded the FAIMS project, led by the University of New South Wales, approximately $950,000 to develop archaeological information management systems between June 2012 and December 2013. Although this project proposal was conceived and initiated in an Australian academic setting, and its primary focus is the production and dissemination of archaeological datasets for research, broader participation is crucial to its success. Most archaeologists in Australia work at consulting firms in the heritage management sector, and they produce the majority of majority of Australian historical and indigenous archaeological data \cite{Ulm2005a}. Consultancies also represent an important industry that is critical to ethical development, particularly with regard to exploitation of Australia’s natural resources. The project also approached state heritage agencies, offering to explore techniques for automating the acquisition and submission of required data by consulting archaeologists and making that data more widely available for research. In an attempt to facilitate consultation with archaeologists who are not directly involved with development, the peak bodies representing the Australian archaeological community of practice were also invited to review and endorse the project. Global innovators in digital archaeology from North America and Europe were invited as well; with this support, we will not have to rediscover approaches, duplicate resources, or develop new standards, but will build on accumulated international experience.

With 40 partners from the university, consulting, and government sectors, drawn from across Australian and overseas[[1]](#footnote-0), and total investment from NeCTAR and co-investment from partner organisations totalling over AUD $2.5m, the FAIMS projects represents one of the most significant international archaeology e-research initiatives undertaken to date.

## Project Goals

The goal of FAIMS is to assemble a comprehensive information system for archaeology. This system will allow data from field and laboratory work to be born digital using mobile devices, processed on the device or exported to local databases, edited and analysed using web applications, and exchanged online through data repositories. Existing standards, components, and tools will be used wherever possible; new ones developed where necessary. When reusing existing resources, open source software will be used wherever possible, and all new development will be open source. Following recent trends in software development, the project will focus on mobile and web applications, and it will employ flexible and reusable data schema rather than requiring users to create custom databases for every project or activity.

Major initiatives of the project include:

* Development of applications for android devices to capture and process data.
* Provision for data export from mobile applications to desktop and web-based tools for processing, analysis, and visualisation.
* Implementation of online repositories for archiving and sharing primary archaeological data produced in Australia or by Australians working overseas.
* Development of new approaches for the production of semantically, as well as technically, compatible datasets.

The end product will be a suite of new tools including a data collection application for mobile devices and an improved online repository that builds from an extant and mature codebase. This system will streamline the management of archaeological data, and it will enhance collaboration, reinterpretation, and comparative study by facilitating the production and dissemination of compatible, high-quality archaeological datasets.

# Stocktaking and Requirements Elaboration

FAIMS invested considerable effort into our stocktaking efforts, and fundamentally revised our project in light of these activities, because we wished to avoid the problems which have undermined other information technology infrastructure projects. As such, we have implemented as fully as possible the tenet that “stocktaking activities are the precondition to designing a successful tool” \cite{Jones2005}. Although our approach and execution was far from perfect in the details, it substantially improved our development plans and offers a model for similar projects.

## Digital Data Survey

In preparation for the Stocktaking Workshop, we circulated a Digital Data Survey to gather information on the use of technology by archaeologists. The survey was aimed at academic, consulting, and government archaeologists or related practitioners who produce or use archaeological or cultural heritage data, including students. The focus of the survey was to describe current patterns of information technology use and define key obstacles to increased use of technology within the archaeological community. While its goal was similar, its scope was smaller and design less complex than the “Strategies for Digital Data Survey” organized by the ADS in Britain to assess the creation, archiving, use and re-use of digital data in archaeology \cite{Condron1999}. Our survey also complements the demographic analyses and vocational training survey of the archaeological community in the UK \cite{Aitchison2003, Aitchison2004} and US \cite{Zeder1997} and an archaeological teaching and learning survey specific to Australia \cite{Ulm2005a}.

We collected 128 responses, 81 of them from Australians, the remaining 39 from overseas participants. Although the sample was smaller and included a higher percentage of academic archaeologists, the profile of FAIMS respondents is largely consistent with that presented by Ulm et al. (\citealt{Ulm2005a}). Due to the nature of survey distribution, combined with an element of self-selection, tech-savvy archaeologists are probably over-represented in our sample. The results nevertheless imply that the initial target audience for FAIMS applications exists today in Australia: a sizeable group of potential users falling on the left half of Rogers’s bell curve as “innovators”, “early adopters”, or “early majority” with regard to the use of technology \cite{Rogers1995}.

Respondents currently employ a range of technologies for their archaeological work, most commonly spreadsheets and word processing software, but often including relational databases and GIS. A majority also use mobile devices. The specific environments proposed by the project (Android and Linux) are, however, less familiar, and will require careful design as well as adequate outreach and training, preferably through tailored online and in-person support. Respondents desire better tools, including mobile device, desktop, and web applications for the creation, analysis, and distribution of archaeological datasets. Most would also like to see greater availability of primary data, are willing to share their own data, and support the idea that researchers should be rewarded for publishing datasets.

## Stocktaking Workshop

In our NeCTAR proposal we requested a major face-to-face Stocktaking Workshop with FAIMS stakeholders prior to the start of development, which NeCTAR accepted as a worthwhile undertaking. The Workshop was held at UNSW on 16-19 August, 2012, and drew about 80 archaeologists, associated researchers, and software developers from Australia and overseas.

Principal topics of the workshop included information management technologies for creating and disseminating high-quality datasets, semantic barriers to producing comparable datasets, and the administrative, legal, and ethical concerns surrounding data sharing. The workshop included plenary sessions providing context for smaller working groups[[2]](#footnote-1). Plenary session speakers discussed current initiatives in digital archaeology (from Australia and overseas), focusing on the challenges to their continued success. Many of the sessions raised issues surrounding data repositories and online publishing, areas less familiar to the Australian audience. Working groups concentrated on information management workflows, field recording needs, and standards. They discussed specific problems, such as mobile device application requirements for various data collection, editing, and sharing, implementation of standards, and issues surrounding sensitive data. The workshop focused on the general characteristics and fundamental requirements of new tools proposed by the project (like mobile device applications for data capture), and specific issues or problems with existing tools that the project would use and adapt (like online repositories).

### Working groups

The working groups formed the heart of the workshop. They included groups that focused on: archaeological survey, excavation, artefact processing (divided between ceramics and lithics), archaeological sciences, federation and data sharing, sensitive data, and strategies for sustaining the tools and services developed by the FAIMS project.

Outputs from the working groups were often different than we had initially expected, but remained extremely valuable. Groups discussing data capture did not produce the detailed technical features we initially expected. Instead, they provided lists of general requirements, summarised in Figure 1. The participants overarching concern was that any applications developed by the project must be extremely flexible, capable of accommodating their existing vocabularies, workflows and practices. Workshop participants also insisted upon the ability to customise user interfaces and data schemata without having to create them from scratch, as is currently necessary when archaeologists use database management systems like MS Access or MySQL.

Throughout the workshop, we saw very little enthusiasm for adopting core data standards. However, many wanted mobile device applications to promote (but not require) data compatibility and other best practices in recording, so long as doing so did not require significant adjustment of existing practice. These expressed requirements, as per Figure 1, forced a radical rethinking of our approach to data capture and standardisation, which we had initially planned to build around a stable (if extensible) core data standards, data schemata, and user interfaces.

Participants reminded us that data can be deemed “sensitive” for a wide variety of reasons: legal, cultural, proprietary, and protective; and that these reasons may change over time, vary by geographical region, and vary between archaeological sub-disciplines. Therefore, online repositories need to be able to authenticate and authorize users to release data only as desired by the submitter.

We were advised to carefully study the sustainability plans of existing archaeological data archives such as Digital Antiquity in the US and the Archaeology Data Service in the UK. We plan to use these plans as well as open source monetization strategies \cite(Kelly2008} to justify providing free access, but to charge for data curation and individual customization of the collection software.

### Focus groups

To maximise the value of the Stocktaking Workshop, self-directed working groups were supplemented by structured focus groups, facilitated by an experienced leader. Focus group topics included:

* Consulting archaeologists;
* Representatives of state heritage agencies;
* Managers of successful overseas online archaeological repositories;
* Field archaeologists (survey and excavation);
* Specialists in artefact processing (lithics and ceramics); and
* Archaeological scientists.

These focus groups were recorded, transcribed and analysed to supplement the information gathered in other fora such as the survey and workgroup discussions.

Focus group responses provided a broader view of archaeological information management, raising concerns about barriers to data sharing (especially inadequate legal and regulatory frameworks), the risks for archaeologists of implementing new technologies, and the sustainability of FAIMS infrastructure. These concerns represent real risks to our project. Bringing the concerns to the forefront compelled us to reconsider the scope and direction of the FAIMS development and to elaborate our sustainability plan. Focus groups (and members of the Steering Committee) also recommended going beyond merely technological development and help improve of Australia’s regulatory framework by assessing the barrier to doing research using datasets held by state agencies, and promoting convergence of reporting requirements around best practices.

## Requirement Elaboration

After the results of the survey and the various Stocktaking Workshop outcomes had been analysed, the FAIMS project entered into a lengthy period of elaboration with our developers, during which a significant portion of the project was reconceived in response to stocktaking activities. Elaboration involved developing user stories for various aspects of the project, turning these stories into technical requirements, and exploring what approaches could be taken to meeting those requirements. Elaboration lasted several months, until November 2012. This lengthy elaboration period allowed us to locate suitable existing software, determine where new development would be needed, and estimate the difficulty and expense of that development. The result was a much better scoped and prioritised project, including detailed software development plan and user requirements. The development plan produced through this lengthy process of stocktaking and elaboration is presented below.

# FAIMS development plan

No information management system has yet been created which allows archaeologists to shepherd their data from digital creation through editing and analysis to online archiving and dissemination. The FAIMS project aims to ensure that the critical parts of such a system are in place, by developing new mobile applications for data capture, enhancing existing data refinement and archiving tools, and ensuring that the various components of the system work together.

We have prioritised the start (digital data capture) and end (online archiving) of the process. We are investing heavily in a customisable mobile device application which supports semi-arbitrary schemas and controlled vocabularies. Both of these aspects balance the production of semantically compatible data and the promotion of best practices on one hand with flexibility and adaptability on the other. The controlled vocabularies cover fundamental archaeological concepts rather than their symbols and, on the model of software internationalisation and localisation, offer translations of the concepts into many different languages of archaeological practice. Using proven data architectures, open data formats, and controlled vocabularies, we can provide well-structured and documented data from our mobile application that is ready for ingest into repositories.

At the other end of the data lifecycle, we are contributing improvements to tDAR <first mention, needs a reference> that will enhance its ability to supply “open linked data:” data that can be located and connected to other data sources in innovative ways \cite{Bizer2009}. We also plan to extend its capacity to store and display the range of data types used by archaeologists, particularly with regard to geospatial data. Taken together, these features will help archaeologists derive more utility from their own and others’ data.

The middle steps of information management - data editing, analysis, and visualisation - represent a mature and complex field, with many commercial and open source tools available to archaeologists. Our principal efforts in this direction involve demonstrating how data can be passed from the mobile device to a robust editing and analysis platform, and later submitted to a repository. Use of open, widely accepted, and well-documented standards throughout the project ensures that the “hooks” are in place to allow the federation of new components at a later date.

For this part of the project we are working with Heurist ([http://HeuristScholar.org](http://heuristscholar.org)) a system developed by Ian Johnson and his team at Arts eResearch, University of Sydney, whose flexible data structures closely align with the methodology used by the FAIMS server. In addition, Heurist already builds and applies schemas through a browser-based GUI, exports those schemas to an XML format (XForms) for use on an Android tablet running ODK (Open Data Kit), and imports collected data back to Heurist, providing a shortcut to building similar schema management capabilities for FAIMS. Heurist mobile collection using these techniques has recently been used at Zagora in Greece and Aquileaia in Italy. For Zagora, extensive legacy data is also collected and managed online within the Heurist database by an internationally distributed research team (Johnson et al. 2012).

Existing, and future, components not built internally by the FAIMS project are critical to our efforts. We are not looking to build a “walled garden”, but hope to establish the foundation of a system that includes options at each stage of data management, many developed externally but federated so as to allow the exchange of data between components. Data created on the mobile device, for example, will be available for export as XML, while tDAR already provides for the import of diverse datasets.

## Design directions

The future of the world is mobile and online. Smartphone sales are climbing rapidly while normal computer sales have stagnated. The search giant Google is turning the majority of its attention to the source of an unexpectedly huge spike of its traffic: the mobile web \cite{Schmidt2011}. Beyond the mobile realm, web applications are also displacing desktop software[[3]](#footnote-2).

In light of these wider trends, our project does not plan to develop traditional “desktop applications”. At the same time, the unique constraints of archaeological operating environments, namely the need to work in remote locations away from the networked world, present unique challenges to mobile- and web-computing paradigms. Many mobile application are web-based, using phones and other devices as clients for software running on servers, and as such require a continuous connection.

While we are making every effort to build a stand-alone mobile application, design constraints, such as core functionality and characteristics of the Android operating system, will require at least an intermittent connection to a local server, usually at the project’s base camp, for activities like configuration, synchronisation, and backup. A continual connection, however, will not be necessary, nor will internet access, allowing the mobile devices to be used for surface survey and other activities even in the most remote locations. Heurist, our application for schema configuration, data editing and analysis, requires a continuous browser connection but this can be over the web or via a local network to the local server (which can be the same server that runs the FAIMS synchronisation server). The repository will be delivered solely online, as ~~it~~ is typical for such services.

## Mobile device applications for digital data creation

The most innovative part of this project involves the development of mobile device applications for the digital collection and management of archaeological data, including spatial data, structured data, free text, images, audio, and video. While digital data collection has clear efficiency advantages \citep[525, 542]{Morgan2012, Hugget2012, Ross2010, Wagtendonk2007}, and modern mobile devices are much more capable than legacy PDAs, no applications specifically tailored for archaeology have yet come into wide use[[4]](#footnote-3). Instead, archaeologists who collect data in the field have tended to spend a great deal of time and effort customising relational database or GIS software running on legacy platforms (particularly MS Access or ESRI ArcGIS and ArcPad). Inspired by these PDA-based implementations of ARCPad \cite{Ross2010, Tripcevich2004, Tripcevich2010, Campana2006}, mobile application developed by FAIMS will exploit the power and screen size of modern mobile devices. Flexible data schemata and user interfaces will be supported, in order to accommodate various archaeological activities, including excavation, survey, and artefact recording - without having to develop schemata and interfaces from scratch. The application will be built for the Android ecosystem, which is more amenable than iOS or Windows to the open source approach promoted by NeCTAR and embraced by this project.

## The mobile device application: A Robust, General Platform for Archaeological Research

The greatest challenge facing developers of archaeological mobile device applications involves providing sufficient flexibility without compromising performance and data integrity - particularly if we seek to improve the compatibility of resulting archaeological datasets. Initially, we planned to produce a data collection application built around stable, core recording standards. The FAIMS Stocktaking Workshop, however, indicated that no consensus has yet emerged regarding core data recording standards for various archaeological phenomena. Participants emphasised that every project or organisation has goals, methods, approaches, and procedures that require fundamental customisation of data recording tools.

We did not, however, want to create an application that was too generic, requiring users to design and build their schema and interface from scratch. Such an approach would too closely duplicate the strengths and challenges of existing mobile GIS applications and relational database management systems. Instead, the approach adopted by our project has been to concentrate mobile device application development on the production of an “interpreter” - a program that can ingest a packet of XML definition documents, using them to generate and deploy a custom data schema and interface. Our decision to build such an interpreter requires more up-front investment than a static (if extensible) data logger would have, but it will produce a much more robust and flexible system. To accommodate its development, we have scaled back some other plans for the mobile device application, but believe we have preserved the core functionality demanded by archaeologists during stocktaking activities.

To customise the application, users can either produce their own XML documents describing a new schema and interface, or retain a developer to produce the documents from existing data models or paper forms; since no application programming would be involved, the time and cost will be contained. Users of the system will be encouraged (but not required) to share their custom definition documents, which will be housed, retrieved, and edited online, producing an ever-growing library. In the future, we envision a web application with a graphic user interface to assist with the development of definition documents, but such a project has been deemed beyond the scope of our current phase of development. After creation of a project using the definition packet, some limited, additional customisation is still possible (e.g., showing, hiding, or renaming fields, revising look-up lists, etc.).

## Semantic compatibility at data creation

Data standardisation, facilitating production of more compatible datasets amenable to comparative and regional studies, is one of the principal goals of the FAIMS project. It became clear from our Stocktaking Workshop that prescriptive approaches that require practitioners conform to predefined core data standards remain unpopular. Our project has therefore focused on non-prescriptive measures, where standards are built into applications and work in the background. Emphasis has also shifted to building semantic interoperability into data at the time of its creation.

First, the FAIMS project plans to develop a library of definition document packets to use with the mobile device application, including schemata and interfaces that accommodate current practice within subdisciplines (e.g., historical or indigenous archaeology), existing workflows of consultancies and academic research projects, state heritage recording requirements, or any combination of these parameters. The library will be made available online through Heurist, with sufficient metadata to aid in the selection of appropriate packages. Heurist will also provide an environment for creating and editing schemas. Any definitions created for this library will be shared across other FAIMS-compliant applications, like the ODK implementation also being developed by Arts eResearch at the University of Sydney.

Schemata provided through the library will, whenever possible, map to controlled conceptual vocabularies in order to promote semantic interoperability of datasets. The interface, however, can employ various aliases for these core concepts, varied to accord with the current practices of different projects, organisations, or archaeological communities. The approach we deploy to accomplish this task builds on the mature and proven internationalisation and localisation concepts standard in the information technology industry \cite{Esselink2000}. In our case, the localisation is not necessarily language or country specific, but related to users’ archaeological practice. The core concept of “stratigraphic unit”, for example, could be mapped to any number of aliases (e.g., “context”, “locus”, “spit”, etc.).

The second major initiative in this direction involves the creation and hosting of the controlled conceptual vocabularies that underlie the process of “Archaeologicalisation” (Arch16n <please no! Regrettabkly I have become quite used to it, and find myself almsot wanting to use it, but I think it is very offputting to our broader public>)[[5]](#footnote-4). These lists will be created descriptively from existing publications, data collection forms, and other aspects of existing practice, with fundamental concepts mapped to various aliases. An investigation conducted by Penny Crook suggests that about 70% of data about any given class of artefact is likely to be ultimately compatible, but will often contain barriers, like divergent terminology, that inhibit automated comparison[[6]](#footnote-5). This tension results in disparate datasets that are difficult for computers to compare, but relatively easy for humans to reconcile manually (either through “loosely coupling” data, or by using an ontology mapping tools)[[7]](#footnote-6). A research sub-project, led by Heather Burke at Flinders University, will focus on extracting concepts and vocabularies from existing practice. The resulting concepts will be hosted by FAIMS and assigned DOIs, allowing their use not only throughout the FAIMS ecosystem, but beyond, much as Open Context now draws taxonomic terminology and ancient Mediterranean place-names from other sources of “open linked data” <<http://opencontext.org/about/concepts>>.

## The Five Files of the Configuration Packet

The configuration packet consists of an XML configuration file plus four definition documents: data schema, user interface, logic, and Archaeologicalisation. These files embody the knowledge representation and the customised user experience required by an archaeologist for a specific project. The configuration packet will be compiled for users’ mobile devices during project set-up by a Linux application server.

The XML configuration file contains project-level metadata, directory structures and file locations, and other project-specific variables, so that one project may deploy these variables across multiple individual instances of the application. For example, project metadata can be shared when a project includes both survey and excavation, and the project director wants to create two separate application instances specific to each activity.

The data schema definition determines what archaeological phenomena and attributes are recorded. The basic approach to data will be to assume that any given archaeological activity (excavation, survey, artefact analysis, etc.) implies a limited set of indivisible recordkeeping entities constituting the “atomic units” of recording. The fundamental entities for excavation, for example, might include “stratigraphic unit”, “special find”, and “artefact group”. These entities are then described by attributes or observations (e.g., composition, dimensions, colours, etc.). Archaeological entities can, furthermore, be grouped. Individual stratigraphic units might be grouped into layers, features, or horizons, which could then be described further by observations pertinent to the whole group. Entities and groups also need to be related to one another in various ways, with the nature of the relationship defined (e.g., above, below, adjacent to, part of, similar to, etc.). The data schema definition defines all such entities, attributes, groups, and relationships.

The user interface and associated logic documents will allow customisation of the interface, including not only layout and local vocabularies, but also a significant degree of procedural logic, automation, and validation. Users will have the usual choices of various input field types (e.g., checkboxes, radio buttons, look-ups, free text fields, etc.), which can be distributed over a number of tabs, each of which will also scroll. The interface skin will inherit Android aesthetics, thereby offering a familiar environment. The logic definition document links procedural code to user interface elements, much like how <in much the same way that> HTML interacts with Javascript. The addition of procedural logic provides for a rich user interface, one which, for example, generates drop-downs as a function of user choices, perform complex validation as a function of other inputs, and makes sophisticated calls to our GIS engine.

The final component of the configuration packet is the Arch16n definition document, which maps local vocabularies used by individual projects against common archaeological concepts. This localisation will use well established i18n techniques, looking for and replacing keyed strings, for example, finding “@StratigraphicUnit” and replacing all instances found with “Context”, in the user interface for that specific project. By replacing strings on display, we can maintain a shared vocabulary independent of, but supporting, different terminologies. By moving concept mapping to the data creation phase of a project, we can produce compatible datasets without the problematic and time consuming ontology mapping at the time of ingest into a repository.

## GIS and Mapping Features

The second major investment of the project into the mobile device application will be the development of mapping and basic GIS features. These features will, at minimum, include the ability to: display georeferenced raster images (e.g., scanned maps and satellite images) in true position; display vector shapes (lines, points, and polygons) in true position; display a user’s position in relation to the raster and vector files; manually draw, select, and edit vector shapes; generate mathematically constrained vector shapes; visualise multiple layers with control over symbology; and manage map projections. In addition, the application will be spatially aware at all times, using on-board GPS or external Bluetooth devices, allowing real-time and automated production of spatial metadata at whatever accuracy is required. We are currently exploring the most effective way to incorporate mapping features into the data entry interface; a tabbed approach similar to ArcPad will serve as a starting point, where an interactive map is displayed on one tab, while additional tabs hold data entry fields.

Most Android (and iOS) mapping software (e.g., Google Maps, Open Street Map) presume a continuous internet connection that can pull maps from online servers. To support fully disconnected mapping, we are using Nutiteq, a commercial software library for GIS rendering on Android devices <[http://www.nutiteq.com/>](http://www.nutiteq.com/). Nutiteq is the only library that we have found that supports necessary GIS operations while operating fully offline. Although it is commercial software, our license with Nutiteq provides for unlimited distribution of their library with our source code, allowing us to keep our open-source approach while providing necessary mapping functionality.

The mapping features of the application will be designed to meet the most common archaeological data entry and visualisation needs during pedestrian survey and other activities that require 3D surface mapping. Visualisation of 3D volumes, which would be of great use for excavations, is beyond the scope of development planned for 2013, but is envisioned for the future. In the meantime, provision will be made to associate an arbitrary number of points defining a 3D volume with a stratigraphic unit record. These points can be entered manually, collected from a GPS (including external dGPS units), or collected from a Total Station. Initially, direct connections will be limited to a selected number of common makes, but provision for import from an SD card during post-processing is also planned. Free or inexpensive, high-quality, third-party CAD Android applications are currently available for projects that require visualisation of three dimensional volumes.

## Data Schema Design

Archaeologists employ different research strategies and approaches. No standard data structure can satisfy such diverse requirements without being so general that it loses utility as a way to organise and manipulate data. As a result, we have chosen to mimic “NoSQL” key-value stores[[8]](#footnote-7). We implement this approach through a Domain-Key Normal Form (DKNF) relational database, as newer NoSQL data stores were <either use ‘are’ here or change ‘implement’ to ‘implemented’ and qualify ‘were’ with some sort of statement about ‘at time of implementation’> either not well supported on Android or were incompatible with GIS software, which generally expects tabular or relational data. After considering the available options available for Android, we chose SQLite <<http://sqlite.org/>> with SpatiaLite spatial extensions <<https://www.gaia-gis.it/fossil/libspatialite/>> as our data store.

Our central data structure is comprised of four tables: ArchEntity, AEntValue, AttributeKey, and Vocbaulary (see Figure 2). However, it is best thought of as a single, virtual table. Imagine a spreadsheet with the top row and the first column locked. <I’m not conveinced that this explanation is clear for the naive reader. I think it would be better to give an example diagram using a row of ArchEntity data, its corresponding AttributeKey rows and values ie somethign more visual with ‘real’ data, perhaps compared with a spreadsheet equivalent>

The unique identifier for an Archaeological Entity resides In the first column, designating a row that will contain all of the information about that entity:

* The “ArchEntity” table acts as the first column <repetitious>, providing an a globally unique identifier (along with identifier-specific metadata<I’d leave this out, maybe put ‘based on the ID of the device in use and the precise time’>) to all information about that archaeological entity.
* The “AttributeKey” table provides a constrained set of attributes concerning the entity<maybe give an example of attributes>. It can be thought of as the frozen row at the top of the sheet with column names representing the things <data?> you want to record about an object.
* The “AEntValue” table represents the data held in the body of the table (each cell being an intersection of an “ArchEntity” row and the “AttributeKey” column; a unique record <you may mean record, but the reader is more likely to think in terms of data value> can be identified by the triple determinant of UUID <most people won’t know what this is>, Timestamp, and Attribute ID).

The relationship between these three tables provides a flexible data schema that does not need to be altered within the database to meet a wide range of use scenarios, and closely parallels the flexible strcuture of Heurist, which is similarly configurable. Only the values contained in the ArchEntity and AttributeKey tables need to be specified in order to define the archaeological entities being recorded along with all of their possible attributes; no aspect of the database structure or associated logic needs to be altered. The final supporting table, “Vocabulary” underpins the Arch16n process that links local terminology to core archaeological concepts.

We intend to never lose data. To achieve that goal, we have instituted an append-only data store, a method modeled on Google’s Protobuffs (Chang et. al. 2008:7). Only writes are allowed to the database; updates and deletions are simulated through inserts. The current records are merely the set of all values with the latest timestamp for a given Universally Unique Identifier (UUID) and Attribute pair. The queries we have designed for normal access to the database will simply disregard all records with an old timestamp or a deleted boolean set to “true”, but the deleted or modified records will remain in the database. This approach makes it trivial to review changes and recover data, as no data is ever deleted or changed.

## 

## User management

In archaeological fieldwork, user identity revolves around the question: “who created this record?”. As such, we will focus on leaving a clear audit trail, identifying any user who changes project data (including, as an option, taking a tiny picture of the entrant every time they add a new record). Combined with the append-only datastore described above, this approach ensures full versioning; changes can be tracked, discussed with the person who made them and, when necessary, reverted - features requested by participants in the Stocktaking Workshop.

## Configuration, synchronisation, and backup

For configuration, synchronisation, and backup, we have implemented a local server, with the mobile devices as clients. The local server runs on a desktop or portable computer running Linux either natively or through a virtual machine on a Mac or PC. Initial setup takes place on the server, with the mobile application pushed out to connected mobile devices. After setup, mobile devices can happily operate independently of a connection to the server, but any time they join the server’s wireless network, they will automatically synchronize their data with one another <is that the case - I thought the sync was now always through the server> as well as the server itself. All records will therefore be available on all devices, and fully up-to-date, whenever connections to the server are present. ~~The server will need to run on its own Linux computer or on a virtual machine on a Mac or PC.~~ We envision stationary activities such as artefact processing taking place with more or less continuous connectivity with concomitant full synchronization, while mobile activities such as pedestrian survey would synchronise upon return to base at the end of the workday. Depending on circumstances, excavation might fall into either category. Data on the server can easily be backed up to to disk or to an online service

The FAIMS server can be accessed and configured through a web interface from any computer, regardless of operating system, on the local network (or Internet if connected). Records can be viewed and edited on the server or on the mobile devices themselves. Users can also manage images and other external files, including linking them to records and producing metadata, generated automatically whenever possible from the mobile device application’s knowledge of time, location, user, etc.

We will allow scans of handwritten notes, sketches, or picture drawings to be associated with individual records; functionality that was stressed as crucial by participants in our Stocktaking Workshop. The server application will also allow the export of data to desktop or web-based software for editing and analysis and, provided no data structures have been altered, re-importation of that data into the mobile system. Finally, the server application will also export data in a format ready for ingest into Heurist and tDAR, providing integration with the project’s data editing and repository platforms.

# Data editing, visualisation, and analysis

Initial project plans included developing a range of analytical and visualisation tools. As we progressed through the elaboration process, however, it became clear that we could not meet all of our initial goals. As a result, we decided to prioritise mobile device development and online archiving and publication. This decision was taken, despite considerable demand for data analysis and visualisation applications from our stakeholders, for three reasons: the transformative potential of modern mobile devices, the importance to the discipline of making primary data available online, and the relative availability of analytical software.

For example, although clear opportunities exist to customise flexible, efficient, and cost-effective open source GIS applications for archaeologists, systems like ArcGIS that are currently in wide use can continue to provide geospatial analysis and visualisation for the time being. Instead of developing rival applications, we are ensuring that data can be exported from the mobile application into ArcGIS and, provided no changes are made to data structures, imported in again. Spatial data will also be exportable as KML, which can then be linked, displayed, and shared via Google Earth.

Another existing resource for archaeologists is Heurist, an online data analysis and visualisation platform developed at the University of Sydney <<http://www.heuristscholar.org/>>. Heurist, a MySQL implementation of a graph database, can be used for routine data editing, but it also facilitates fundamental reorganisation and rethinking of data. For example, its visualization and association capabilities help decisions regarding which excavation contexts should be included in which horizons, or which individual artefacts should be classified as which types. FAIMS is working with Arts eResearch to facilitate automated data transfer from the FAIMS mobile application, through Heurist, to tDAR. Under this workflow, data produced on the mobile device application can be passed to Heurist for further analysis and restructuring before being submitted to an online repository. Data and visualisations can also be exposed directly for publication from Heurist, as has been done with the Dictionary of Sydney <<http://home.dictionaryofsydney.org/>>.

Taken together, existing tools provide a number of ways for archaeologists to edit, analyse, and visualise data created using our mobile application (or other legacy data) and export the results to an online repository.

# Online archiving and publication: an Australian Archaeological Database

Online archiving and publishing are the most mature realms of the project. FAIMS plans to build on the foundations laid by the Australian Historical Archaeology Database (AHAD) <<http://www.ahad.edu.au>>, maintaining and improving its Australian implementation of the Digital Archaeological Record (tDAR) <<http://www.tdar.org>> for use as a stable, long-term repository for archaeological data.

The AHAD implementation of tDAR was selected by FAIMS as its repository of first resort. tDAR is an open-source, online platform for archiving archaeological datasets and associated reports, images, vocabularies and other records. It is built upon a PostgreSQL object-relational database and includes intuitive interfaces enabling archaeologists to upload and manage their own data and create appropriate metadata, without specialised training in database management or archival services. It can accommodate datasets that vary considerably, and includes provisions for mapping project ontologies and vocabularies to its internal schemata, thus creating compatible datasets for integrated analysis \cite{Speilmann2011}. In addition to copying structured data into its internal tables, tDAR stores imported datasets in their native format. tDAR supports a comprehensive metadata schema to describe uploaded resources, including the allocation of one or more keywords for cultural groups, geographic locations, temporal phases, material classifications, investigation types, participants, and sponsors. tDAR also has controls to protect sensitive data.

Instead of being limited to historical archaeology, the revised implementation of AHAD will include indigenous and maritime archaeology, as well as data generated by Australian archaeologists working overseas. The database will be renamed to reflect its increased scope. We are coordinating all development with tDAR, inheriting their changes and contributing our development work.

Our project will invest in adding features to the repository, including:

1. Automated data import from the mobile device application and from Heurist;
2. implementation of UUIDs allowing row-level data access from the web;
3. spatial data storage and enhanced mapping functionality;
4. improved management of sensitive data; and
5. integration with Australian infrastructure, including Australian Access Federation support and NeCTAR Research Cloud / Research Data Storage Infrastructure hosting.

FAIMS will ensure that data, schemata, and controlled vocabularies can be exported from our mobile device application into the repository as easily and automatically as possible, translated into a format consistent with the object-relational model of tDAR. Any record groups, cross-references, and links created in the field will be preserved, as will all metadata associated with a record.

Our project also plans to improve the granularity of data in the repository by consistently implementing Digital Object Identifiers (DOIs) <<http://www.doi.org>> as citeable universally unique identifiers for every entity in the databases. DOIs will be minted at the time of ingest into the repository, but based on unique identifiers assigned by the mobile application during data creation. Unique identifiers will help move the repository towards row-level pages - a “one page per pot-sherd” model - where entities in the database can be addressed, viewed, annotated, and exported individually, facilitating the retrieval, manipulation, linking, recombination, and repurposing of data \cite{Kansa2010f, Kansa2010c}.

After automated import and improved data granularity, enhanced mapping functionality is our third priority for improvements to the repository. Users should be able to store, search, and display vector data using a base map or satellite image as a background. Doing so will require adding support to tDAR for geographic data types: KML, shapefiles, or through the use of spatial data types implemented by PostGIS, a PostgreSQL extension. <<http://postgis.refractions.net/>>. Visualisation will be provided by improving tDAR’s existing deployment of the Google Maps Engine, which currently only contains a single location for each project drawn from project-level metadata. The enhanced implementation will allow visualisation of selected record-level geospatial data and selection of records using a map interface (either through the selection of a region or jurisdiction, or by drawing a polygon).

The fourth set of new features proposed for the repository involves improved management of sensitive data. tDAR is already capable of obfuscating locations, a feature that is being refined. Furthermore, over the course of the FAIMS workshop it became clear that users need a way to hide or release specific data according to a number of parameters (data ownership, cultural sensitivity, pre-publication embargoes, etc). As such, we plan improvements that will allow data to be hidden from selected users according to various criteria. Part of this solution will involve joining the Australian Access Federation <<http://www.aaf.edu.au/>> as the basis of a comprehensive user identification and authorisation system. Unencumbered data could still be accessible without a login, as could useful derivative information that does not expose underlying sensitive data (for an example, see “The Aboriginal Sites Decision Support Tool” <http://www.environment.nsw.gov.au/licences/AboriginalSitesDecisionSupportTool.htm>).

AHAD is currently hosted by the La Trobe University library, but the improved repository will be implemented on the NeCTAR Research Cloud <<http://www.nectar.org.au/research-cloud>> and the Research Data Storage Infrastructure (RDSI) <<http://rdsi.uq.edu.au/>>. As dynamic scaling will not be necessary, this implementation will primarily involve migration to stable virtual machines on these infrastructures.

With these improvements to the tDAR/AHAD repository, FAIMS will ensure that a stable, long-term archive is a available for Australian archaeological data, one which will promote data discoverability and portability. The result will be a robust system for data curation and dissemination.

## Dataset publication

The final initiative undertaken by FAIMS involves establishing appropriate dataset publication outlets for the Australian archaeological community. We intend continue the work begun by NSW Archaeology On-Line <[http://nswaol.library.usyd.edu.au/](http://nswaol.library.usyd.edu.au/index.jsp?page=home)>, a “gray literature” archive established by Martin Gibbs and Sarah Colley at the University of Sydney (Gibbs and Colley 2012). Going forward, consultants’ reports will continue to be solicited for scanning, digitisation, and distribution online with appropriate metadata, but archived in the project repository and expanded to include reports from across Australia. A similar initiative by the Archaeology Data Service in the UK has been enormously successful, so much so that Julian Richards, Director of the ADS, now insists that “the grey literature is the white literature” in the UK[[9]](#footnote-8).

We have also entered into discussion with the *Journal of Open Archaeology Data* <<http://openarchaeologydata.metajnl.com/>> to accept datasets housed in our repository (tDAR is already an accepted repository). *JOAD* functions as a tradition (online) journal, with a peer review process, publishing substantial datasets with accompanying contextual information and discussion. In the future, it may be worthwhile to establish an independent dataset journal on the model of the *JOAD* catering specifically to Australian and Australian-generated data. Such a journal could, for example, publish supplemental data supporting articles in Australian journals, as well as stand-alone datasets.

Publishing datasets through an online journal would not only make such datasets more widely available, and would also provide a professional incentive for university-based archaeologists (as well as their schools and universities, who typically receive HERDC funding for articles published by academic staff). FAIMS is assembling example datasets to demonstrate the potential of such online publication.

# Sustainability

A focus group of international participants in our project identified sustainability as the greatest short-term challenge facing the project. Currently, the FAIMS project is funded only through the end of 2013, with co-contributions from Development Organisations ensuring maintenance through the end of 2014. Plans for short- and long-term sustainability, however, are being produced based on recommendations from project participants who are involved with similar, successful initiatives overseas such as the Archaeology Data Service in the UK (Kintigh 2006).

Short-term plans for funding the project in 2014-2015 involve applying for further infrastructure grants and charging mobile application customisation and repository data ingest fees. Grant application efforts will focus on the ARC Linkage infrastructure, Equipment, and Facilities (LIEF) program. Smaller, but still substantial, grants are available through e-research initiatives at participating Australian universities and through the Google Research Awards program.

We are requesting ~~the~~ that archaeologists based at Australian universities include a line item for data management and archiving in their ARC Discovery and Linkage applications, with the amount required varying by the size and complexity of the dataset. Using examples from overseas repositories, we anticipate repository archiving fees to range from a few hundred dollars to perhaps $10,000 for the largest and most complex projects. At least initially, archiving fees will include assistance with mobile device implementation, since our interpreter-based system is relatively easy and inexpensive to customise, and will produce well-formed data ready for submission to tDAR/AHAD, reducing the costs of ingest. At the same time, we hope to demonstrate to consultancies that the mobile application will produce sufficient efficiency gains to warrant payment for an annual service contract that would include mobile application customisation and data archiving.

A longer-term sustainability plan, for the years 2015-2020, is also being formulated[[10]](#footnote-9). The outlines of this plan include creating a university-based research centre or a not-for-profit organisation with a permanent staff as an institutional home for the project. We will continue to raise money through short-term grants and customisation and ingest fees, but hope to supplement that income by seeking longer-term funding from private foundations and other sources.

We hope that closer collaboration with state heritage agencies will increase the desirability of the system, by providing automated submission of legally required information, and that the usefulness of online datasets will make their creation more worthwhile through increased citation and other incentives. It is not inconceivable that the ARC or Australian-based archaeological journals will follow the lead of their European and American counterparts and begin requiring online availability of datasets as a condition of grant awards or publication, a service FAIMS can provide[[11]](#footnote-10). Disciplinary changes appear to be coming that will promote or require open, online publication of research data, a development we are prepared to help Australian projects and organisations accommodate.

# Conclusion

The FAIMS project is building information infrastructure which incorporates mechanisms for the production and dissemination of high-quality and reusable datasets into the entire archaeological data lifecycle. We have consulted extensively to ensure that the tools we produce will meet the varied needs of archaeologists and related professionals, and that approaches to data standardisation and compatibility are acceptable to practitioners. Surveying the archaeological IT landscape, the project chose to focus on developing new mobile applications for data collection, improving an existing online repository, and ensuring that these and other new and existing components could work together in a modular, federated system. The mobile device applications under construction centre on an interpreter that can produce highly customisable data schemata and user interfaces from XML definition documents. At the same time, techniques from localisation and internationalisation are being deployed to map the vocabularies of individual projects to core concepts from project startup.

Our mobile applications incorporate mapping, versioning, and other key features articulated by archaeologists. Data will be exportable from the mobile application ready for ingest into an implementation of tDAR for archiving, or it can first be exported to a range of existing platforms for editing and analysis. tDAR itself is being enhanced so improve features such as mapping and the management of sensitive data, and improved so that it can expose data in a more granular form, facilitating its location, extraction, reuse, and analysis. This holistic approach to the archaeological data lifecycle, and our extensive consultation with archaeologists, will produce tools that are more likely to actually see use in the field and laboratory, and therefore to address long-standing problems with data availability that currently hinder archaeological research.

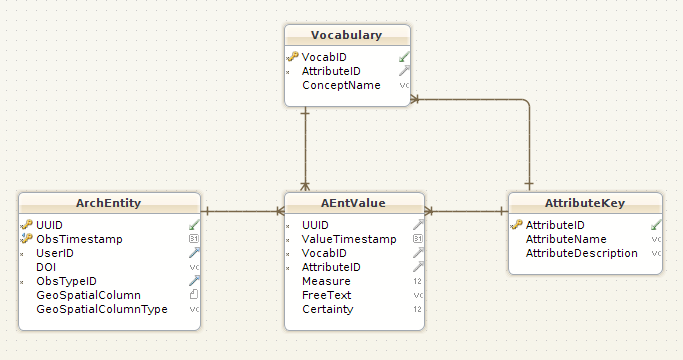
# 

# Figures

## Figure 1: Mobile device requirements from the Stocktaking Workshop

|  |
| --- |
| * “it needs to be as easy as paper recording”, * “accommodate GIS input, output and visualisation of highest possible precision”, * “it needs to facilitate data streamlining and cleaning”, * “we need to be able to take paper notes, sketches, add files and annotate them”, * “we want to control data input, and impose vocabularies”, * “increase commensurability and consistency” * “save, reuse, customize, proliferate schemas” * “work without web connectivity” * “synchronize data in the field” * “collect contextual and user data, such as weather conditions, identity of user, etc.” * “have autofill functions for dealing with repetitive data” * “indicate the ‘level of confidence’ or ‘level of subjectivity’ that the user/data entry person has with respect to any of the fields” * capture changes when you upgrade/change records or schema (versioning), track to user * be able to re-label columns (aliasing) * “be really rugged to handle Australian extremes (dirt, rain, heat, sun glare etc)” * “have a LONG battery life” * “has to be affordable” |

## Figure 2: Mobile device database entity relationship diagram



1. For a complete list of participating organizations, see: <http://www.fedarch.org/wordpress/?page\_id=162> [↑](#footnote-ref-0)
2. Workshop documentation including the keynote video, plenary session slides and audio files, discussion group quidelines and summary audio files are available on the FAIMS webpage <<http://www.fedarch.org/wordpress/?cat=10>>. [↑](#footnote-ref-1)
3. For a discussion of the advantages of web applications, see the prescient and still-useful 2001 essay by Paul Graham, “The Other Road Ahead: Web-based software offers the biggest opportunity since the arrival of the microcomputer” <<http://www.paulgraham.com/road.html>>. [↑](#footnote-ref-2)
4. a variety of mobile applications have been produced by specific projects for their research needs, to name a few the Integrated Archaeological Database <<http://www.iadb.co.uk>>, Cybertracker <[www.cybertracker.org](http://www.cybertracker.org)>, or the iPAD application used by the PARP: Porta Stabia project of the University of Cincinnati <<http://classics.uc.edu/pompeii/index.php/project.html>>. [↑](#footnote-ref-3)
5. Arch16n, or Archaeologicalisation, is a numeronym coined using the model of other localisation standards like i18n, a shortened form of internationalisation, as pioneered by DEC \cite{Texin2010} in the 1970s. [↑](#footnote-ref-4)
6. Part of ongoing research, presented at the FAIMS Workshop. Slides and Audio may be found: <<http://www.fedarch.org/wordpress/?p=282>> [↑](#footnote-ref-5)
7. See, for example, (Kansa and Bissel 2010) for “loosely coupled” data. [↑](#footnote-ref-6)
8. Native XML stores (like OCHRE;<http://ochre.uchicago.edu/page/xml-database-structure>), graph databases like CouchDB (used by OpenDig; <http://www.opendig.org/>), and Google protocol buffers (<http://code.google.com/p/protobuf/>) were considered. For an introduction to NoSQL, see “NoSQL” (<http://en.wikipedia.org/wiki/NoSQL>) and “NOSQL Databases” ( <http://nosql-database.org/>). [↑](#footnote-ref-7)
9. Audio and Slides from his presentation available <<http://www.fedarch.org/wordpress/?p=284>> [↑](#footnote-ref-8)
10. [Concerns about long-term sustainability of fast-sprung digital initiatives have tempered adoption and retain high press in literature: Jeffrey, S. (2012). A new Digital Dark Age ? Collaborative web tools , social media and long-term preservation. World Archaeology, 44(December), 553–570.](http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp) [↑](#footnote-ref-9)
11. See, for example, the data management requirments of the US National Science Foundation <<http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp>>. [↑](#footnote-ref-10)