### FAIMS Development Plan Narrative - Addendum to the Software Development Requirements for the Steering Committee

28 October 2012

### Digital data creation: mobile device applications

The most innovative - and risky - part of this project involves the development of mobile device applications for the digital collection of archaeological data. While digital data collection has clear efficiency advantages, and modern mobile devices are much more capable than legacy PDAs, no applications specifically tailored for archaeology have yet come into wide use [cf. [www.IADB.org.uk](http://www.iadb.org.uk/) or<http://archeolink.org/UK_index.html>]. Instead, archaeologists who collect data in the field still tend to spend a great deal of effort customising GIS software running on legacy platforms. Inspired by these older PDA-based implementations of ESRI ARCPad [references: Ross 2009 Wernke & Tripcevich 2010], the applications developed by FAIMS will exploit the power and screen size of modern mobile devices. Three components are planned: excavation, survey, and artefact recording, all 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).

Providing sufficient flexibility without compromising performance and data integrity constitutes the greatest challenge facing developers of mobile device applications for archaeology. Initially, we had planned to solicit a certain degree of convergence from the Australian community of practice (an undertaking that would also promote semantic interoperability) and produce a relatively static, but robust, data collection application - one with a stable core (in terms of the data schema and interface) that could be extended and, to an extent, customised. One outcome of the FAIMS workshop was a clear sense that archaeologists were only willing to compromise so far on this point, and that the “core data recording standards” needed to underpin a stable and robust mobile database was likely out of reach, even if it was designed to be extensible. It became clear that every project or organisation has goals, methodologies, and procedures that require fundamental customisation of data recording processes.

We did not, however, want to create an application that was too generic, requiring users to build their schema and interface from scratch - such an approach would too closely duplicate the strengths and challenges of existing mobile GIS applications (e.g., GIS Pro, ARCPad, etc.) and desktop or client-server databases (Access, SQL Server, MySQL, etc.).

Instead, the approach adopted by our project has been to concentrate our mobile device application development on the production of an “interpreter” - a program that can ingest, e.g., an XML document and use it to generate a custom data schema and interface. We will also produce a library of these definition documents, including schemata based on jurisdictional requirements, the practices of subdisciplines (e.g., historical or indigenous archaeology), the current practices and workflows of consultancies and academic research projects, or any combination of these parameters. Schemata provided through the library will, whenever possible, map to controlled vocabularies, while the interface can employ various aliases for these core concepts, varied to accord with current practice amongst different archaeological communities. For example, the core concept of “stratigraphic unit” could be mapped to any number of aliases (context, locus, spit, etc.). Some limited customisation (showing / hiding fields, renaming fields, revising look-up lists) will be allowed on the device. Users requiring further customisation (or entirely new definition documents) could either produce their own XML documents describing a new schema and interface or hire FAIMS (or another developer) to produce the document from existing data models or paper forms; since no programming would be involved, the time and cost would be contained. Users of the system would be encouraged (but not required) to share their definition documents, which could be housed, retrieved, and edited through an online system like Heurist (<http://heuristscholar.org/heurist/help/tour.html>), producing an ever-growing library, In 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.

Data schemata produced by the definition document would determine what archaeological phenomena and attributes would be recorded. The basic approach to data will be to assume that any given archaeological activity (excavation, survey, artefact analysis, etc.) has some sort of indivisible entity constituting an “atomic unit” of recording (stratigraphic unit, survey unit, artefact, etc.) that needs to be described by a series of observations. These entities can, furthermore, be grouped arbitrarily in various ways. Individual stratigraphic units, for example, might be grouped into layers, features, or sites), which could then be described further by observations pertinent to the whole group. Units 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, etc.). Within this broad and flexible framework, individual projects or organisations will have the freedom to design schemata to meet their needs.

The definition document will also allow customisation of interfaces, including not only layout, but also the use of controlled vocabularies, and a degree of automation and validation. Users will have the usual choices of various input field types (check boxes, radio buttons, look-ups, text fields, etc.), which can be distributed over a number of tabs, each of which will also scroll. Interfaces will be built according to current Android standards, offering a familiar environment. Definition documents will also generate Javascript, allowing automation (such as importing data from previous entries, setting defaults, populating drop-downs with controlled vocabularies) and validation (such as ensuring completeness and adherence to simple rules). In keeping with our commitment to deploying simple and widely-used standards possible, we plan to implement the XForms Model, a device-independent XML form definition (<http://www.w3.org/MarkUp/Forms/>).

The second major investment of the project into the mobile device application will be the development of mapping features. These features will, at minimum, include the ability to: display georeferenced raster images (e.g., maps and satellite images) in true position; display vector shapes (lines, points, and polygons); display one’s position in relation to the raster and vector files; generate mathematically constrained vector shapes; manually draw, select, and edit vector shapes; visualise multiple layers with control over symbology; 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. We are currently exploring the best way for the interface to unite mapping with data entry, but 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).

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 most likely 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 volume with a stratigraphic unit record; such points can be entered manually, collected from a GPS (including external dGPS units), or collected from a Total Station (initially, support will be limited to a selected number of common makes, but provision for import from an SD card during post-processing is also envisioned).

The FAIMS project has committed to developing open-source software (OSS) and to writing as little new code as possible. Together with Intersect, we are evaluating available open source software libraries for both data collection and GIS / mapping. Mobile data collection software under consideration includes:

* Open Data Kit ([http://opendatakit.org/](http://opendatakit.org/),))
* OpenXData and mForms2 ([http://www.openxdata.org/category/org/](http://www.openxdata.org/category/org/).))
* JavaRosa (<http://www.dimagi.com/javarosa/>)
* Epicollect (<http://www.epicollect.net/>).

No suitable open source mobile GIS or mapping software has yet been located, but several mature GIS and CAD OSS libraries exist, including:

* GeoTools (<http://geotools.org/>) and other software from the Open Source Geospatial Foundation (http://www.osgeo.org).
* Open Cascade (<http://www.opencascade.org/>)
* BRL-CAD (<http://brlcad.org/>)
* OpenSCAD (<http://www.openscad.org/>)

In the OSS realm, CAD software is more varied and mature than GIS software. CAD software can serve as the engine for GIS (see MicroSurvey and Manifold for examples of commercial software that use this architecture:<http://www.microsurvey.com/> and http://www.manifold.net). In addition, CAD software managed 3D volumes (most GIS software is limited to 3D surfaces), which could, potentially, allow for a single spatial solution to both mapping and manipulating the 3D volumes produced by excavation (a feature we would like to add in a future stage of development). Finally, we are still evaluating options for a database management system; many OSS DBMSes are available...

Aside from the interpreter-based data collection tool with mapping functionality, a limited range of other essential features has been identified. The application will know its user (i.e., require login) and store project-level information, allowing the automatic production of important record-level metadata. it will function completely offline for use in remote locations. A means of synchronising all devices in use will be provided (optimally, devices will automatically form a “mesh” when they are within wi-fi range of one another and synchronise in that manner; if such an approach encounters technical obstacles, synchronisation will be done via USB). Users will have the ability to manage images and other files, including linking them to records and producing metadata (generated automatically whenever possible from the application’s knowledge of its location, user, etc.). Note that file management will allow incorporation of scanned handwritten notes and sketches or technical drawings into the system (with appropriate metadata and record links), functionality that was stressed as crucial by participants in our Stocktaking Workshop. The application will 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 application will export data in a format ready for ingest into tDAR/AHAD and Open Context, providing integration with the project’s repository and publications platform.

A system that uses an interpreter to produce custom data schemata and interfaces from simple-to-edit XML files will require more up-front investment than a static (if extensible) data logger, but it will produce a more robust and flexible option for archaeologists. To accommodate its development, we have scaled back some other plans for the mobile device application, but believe we can still ensure necessary core functionality, including: (1) creation, searching, and editing of records and groups; (2) mapping functionality, including georeferenced image display and vector file display, creation, and editing; (3) image and file management, including metadata production; (4) export of well-formed data to editing and analysis platforms and for online archiving.

### Online archiving and publication: an Australian Archaeological Database

Online archiving and publishing are the most mature realms of the project. The project’s approach has two components. First, FAIMS plans to build on the foundations laid by the Australian Historical Archaeology Database (AHAD), maintaining and improving an Australian implementation of the Digital Archaeological Record (tDAR) for use as a stable, long-term repository for archaeological data. Second, FAIMS seeks to facilitate the online publication of archaeological data through an Australian implementation of Open Context, a platform optimised for the dissemination (as opposed to archiving) or data. The two underlying databases - tDAR and Open Context - are complementary components of our approach to online archiving and publication.

The core of our online archive will be an enhanced version of the Australian Historical Archaeology Database (AHAD). Instead of being limited to historical archaeology, the revised implementation of tDAR/AHAD will include indigenous archaeology and data generated by Australian archaeologists working overseas. The database will be renamed from AHAD to “the Australian Archaeological Database (AAD)” or something similar to reflect its increased scope. The underlying software will remain an asynchronous mirror of the tDAR code (not content) hosted by Arizona State University. We are taking pains not to “fork” the code base; we are coordinating with tDAR to inherit all of the changes they make, and to feed all of our development work (performed by VeRSI, the Victoria eResearch Strategic Initiative) back to tDAR. As noted above, tDAR/AHAD can accommodate datasets that vary considerably, although it does include provisions for mapping ontologies and vocabularies when data is ingested, in order to help produce more compatable datasets.

FAIMS will invest its resources adding features to the tDAR/AHAD. The broad categories of improvement include, in order of priority: (1) enhanced and, to the degree possible, automated data import and export; (2) GIS / mapping functionality; (3) improved handling of sensitive data; (4) data editing and other features.

To improve data portability, each entity in the database needs a persistent, globally unique identifier. Our project plans to improve the granularity of data in tDAR/AHAD by consistently implementing Digital Object identifiers (DOIs;<http://www.doi.org/hb.html>). DOIs will permanently and uniquely identify all entities, not only within any given dataset, but across all datasets in tDAR/AHAD, and indeed across all data sources that use DOIs. DOIs will help move tDAR/AHAD towards a “one page per pot-shard” model, where individual entities (artefacts, excavation contexts, survey units, projects, people, documents, etc.) can be addressed, viewed, annotated, and exported individually. If, for example, a search returns results from many datasets, those results can be exported as a new dataset, with the assurance that each entity in it will have a unique identifier. DOIs can incorporate existing identifiers, and they can be aliased like URLs to be more readily accessible to people. DOIs facilitate the retrieval, manipulation, linking, recombination, and re-purposing of data, and constitute the most fundamental upgrade to tDAR/AHAD.

Improved granularity of data will aid the export and reuse of data, but other features also contribute towards data portability. While tDAR/AHAD represents a repository optimised for long-term data storage, Open Context serves as an example of a publication service designed to exchanging data using widely adopted standards. The features of Open Context that compare favourably with tDAR/AHAD include faceted navigation (which allows users to explore hierarchical data with a greater certainty that search results will be comprehensive), and export of data as XML using the Atom syndication format (widely used standards that can communicate complex data structures) ([http://www.opencontext.org/about/](http://www.opencontext.org/about/).)). Since it is open source, an Australian implementation of Open Context is feasible, and negotiations are underway with its administrators to establish an Australian mirror similar to the arangement between tDAR and AHAD. With improved granularity and implementation of DOIs in tDAR/AHAD, interoperability with Open Context will be possible. The required extract-transform-load processes will be written internally at UNSW, yielding a system where data can be archived in tDAR/AHAD, and published through Open Context - an arrangement that will likely be less costly in time and resources than adding the features of Open Context to tDAR/AHAD. Open Context could also serve as the technical basis of an Australian equivalent to the *Journal of Open Archaeological Data* (see below), and more generally provide an example of an alternative end-point for data in the FAIMS ecosystem.

We also want to ensure that data can be imported as easily as possible, whether from mobile device applications or legacy sources. Data, schemata, and controlled vocabularies from mobile device applications will import automatically, translated into a format consistent with the object-relational model of tDAR/AHAD. Any record groups, cross-references, and links created in the field also need to be preserved, as does all metadata associated with a record. DOIs, assigned at the time of data creation, will allow the maintenance of metadata and internal references through the import process. For legacy data imported from other sources than FAIMS-compliant mobile device applications, DOIs will be assigned upon import, and new tools to streamline the creation of metadata will supplement the ontology mapping features that are already part of tDAR/AHAD. Finally, we plan to expand the data types supported by tDAR/AHAD to include audio and video files.

After data import and export, mapping functionality is our second priority for improvements to tDAR/AHAD. The repository should be able to store, search, and display geographic vector data using a base map or satellite image as a background. Doing so would require adding support for geographic data types (KML, shape files, or through the use of PostGIS, a PostgreSQL spatial extension;<http://postgis.refractions.net/>), and providing some means of visualisation. Two approaches to the latter are being considers: (1) the integration of a OSS GIS or CAD library with tDAR/AHAD, or (2) integration with a web service such as Google Earth Pro. Either approach should allow visualisation of selected records in three dimensions and selection of records using a map interface (either through the selection of a region or jurisdiction, or by drawing a polygon). As discussed above, implementation of a CAD library for mapping would also allow the display and manipulation of 3D volumes (in addition to surfaces), a feature that may eventually pay dividends for visualising excavation contexts (a task that GIS-based mapping software does not handle so well),

The third set of new features proposed for tDAR/AHAD involves improved management of sensitive data. tDAR is already capable of obfuscating locations. Currently, tDAR randomises the location of an object within a 1 sq km area, which sometimes produces odd results (sites located in the ocean, etc.). As a result of these problems, the feature has been disabled in AHAD, and a more nuanced and adjustable (at the discretion of the dataset owner) way of obscuring the location of sensitive archaeological features is desirable. 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, etc). As such, we plan improvements that will allow data to be hidden from some 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 data that did not expose underlying sensitive data (for an example, see [link to NSW OEH maps presented at the workshop]).

The fourth and final new addition planned for tDAR/AHAD involves improving the ability of contributors to edit their data. As a repository, tDAR/AHAD currently expects to ingest relatively “clean” data, but is also in the process of developing an editing interface based on the Open Data (OData) protocol, which allows querying and updating data over Web (<http://www.odata.org/>). FAIMS will contribute towards development of this feature. Likewise, a few improvements to search capabilities and other features have also been proposed, mostly in support of efforts already underway at tDAR. FAIMS is actively coordinating with tDAR (a representative from tDAR - as well as one from Open Context - sits on our Steering Committee to ensure that efforts are not duplicated)

.

With these improvements to tDAR/AHAD, FAIMS will ensure that a stable, long-term archive is a available for Australian archaeological data, one which will use persistent, granular identifiers and comprehensive, standardised metadata to promote data discoverability and portability. Data produced by FAIMS-compliant mobile device applications will import seamlessly into tDAR/AHAD, while tools for metadata creation will ensure that legacy data is discoverable. Interoperability with an Australian implementation of Open Context will add additional online publication options, included faceted search and export using standard web protocols, further encouraging reuse and re-purposing of data. Storage and visualisation of geographic data, including geography-based searching, will be added to this system, along with features to protect sensitive data and allow some editing within the system. The end result will be a robust system for data curation and dissemination.

### Data editing, visualisation, and analysis

Initial project plans included developing a range of analytical tools for archaeology, including a customised OSS desktop geodatabase that could safely exchange data with the mobile device application (without altering data structures), and an experimental data warehouse for testing the applicability to archaeology of new knowledge discovery techniques developed in other fields. As we progressed through software development planning, however, it became clear that we could not meet all of our initial goals, especially considering NeCTAR’s requirement (re-emphasised in their project revision request) that all software developed must be production ready by the end of 2013; no prototypes or beta versions are acceptable. As a result, we decided to prioritise mobile device development and online archiving and publication. This decision was taken because of (1) the transformative potential of modern mobile devices, and (2) the importance to the discipline of making primary data available online (combined with the opportunities presented by developing partnerships with ongoing initiatives in Australia and overseas). Furthermore, it was determined that existing analysis and visualisation tools could meet many archaeologists’ needs - although this area is one we plan to revisit in a later stage of development.

According to our survey (and personal experiences), the technology most familiar to archaeologists is desktop GIS, especially ESRI ARCGIS. ARCGIS is a complex (and sometimes erratic) program that includes many features archaeologists never use. It is also expensive to licence, especially when the additional costs of Spatial Analyst and other useful extensions are considered. It would, therefore, be desirable to customise, an OSS GIS such as, for example, GRASS, for archaeological use. Such a customisation could streamline the interface for typical archaeological work-flows, and automate geospatial analyses frequently used by archaeologists. The implementation could also be done in a way to disallow changes to data structures, thus ensuring that data could be passed back and forth between the mobile application and a desktop GIS (or, more ambitiously, data structures could be modified and then synchronised with the mobile application). Clear opportunities exist to produce a more flexible, efficient, and cost-effective OSS desktop GIS for archaeologists, but systems like ARCGIS that are currently in wide use can continue to fill this role now, without endagering the overall goals of the project. As a result, we are ensuring that data can be imported from the mobile application into ARCGIS and, provided no changes are made to data structures, exported back out.

Google Earth is another spatial visualisation tool that will be familiar to many archaeologists. Spatial data recorded in the FAIMS mobile device application will be exportable as KML, which can then be linked to and displayed on Google Earth. We are also exploring the addition of a Google Earth plug-in to tDAR/AHAD so that users can view these visualisations without leaving the repository website.

Another existing resource for archaeologists is Heurist, an online data analysis and visualisation platform developed by Arts eResearch at the University of Sydney (<http://www.heuristscholar.org/>). Heurist, a MySQL implementation of a graph database, facilitates fundamental reorganisation and rethinking of data (as, for example, decisions regarding which excavation contexts should be included in which layers, or which individual ceramics should be categorised as part of which types). Heurist also includes a number of useful spatial, temporal, and conceptual visualisation features. After data is produced by the mobile device application, projects wishing to undertake further analysis and restructuring of data will be able to pass it through Heurist before submitting it to tDAR/AHAD or another online repository. Data and visualisations can also be exposed for publication, as has been done with the Dictionary of Sydney, which is powered by Heurist (<http://home.dictionaryofsydney.org/>).

Taken together, these existing data analysis and visualisation tools will provide a number of ways for archaeologists to analyse and visualise their data, as well as publish the results online.