

## IDMB Archaeology Case Study: Summary

### Context

Archaeology was chosen as a discipline case study for IDMB as part of the research data practice data-gathering, piloting of training and of technical data management solutions. These three components were interlinked and involved the same project staff. The technical pilot began with a consideration of metadata needs and strategies, and then continued with trial implementations using EPrints and SharePoint 2010. Archaeology provides a rich set of case studies because:

- It mixes humanities and science research practices;
- Includes very data-intensive research practices, producing large numbers of data objects, big data assets, and complex data transformations;
- Commonly develops very complex i.e. 'short fat' rather than 'tall thin' data structures;
- Researchers have broad disciplinary backgrounds and hence experience of data management.

The archaeology department at the University of Southampton includes more than twenty FTE academic staff, approximately half this number of research and technical staff, and has a significant cohort of postgraduate research students. Of the latter approximately twenty are currently focussed on archaeological computation and can be considered an expert group within the larger sample in terms of this case study. This group provided a broad range of specialist information, largely focussed on specific data types and methods such as laser scanning, geophysics and linked data. In addition to the online questionnaire and structured interviews outlined in the IDMB survey report. In the interim report we conducted a large number of follow-up and additional informal interviews in order to define prototype metadata strategies and technical infrastructures. The staff and research students involved in this case study exhibited a broad range of data management needs and expertise. In the workshops we specifically targeted data-intensive users with limited formal training or knowledge of data management expertise and this formed the basis for informing the pilots. As far as possible these were therefore driven by perceived need and current practice rather than idealised behaviour requiring significant changes for the researchers involved.

The IDMB archaeology case study was able to draw on recent expertise gained on a number of funded and unfunded research projects. These include:

- The AHRC Portus Project, which has for four years been developing and critiquing computational methods in archaeological practice, with data management based around the Archaeological Recording Kit (ARK). In particular it has built upon recent advances in geophysics, computerized excavation recording, topographic planning, laser scanning of buildings, and detailed object recording, in order to develop an integrated approach to three-dimensional recording, interpretation, management and dissemination;
- The AHRC Reflectance Transformation Imaging for Ancient Document Artefacts project, which has developed a number of workflow and data management approaches to a specific range of archaeological imaging technologies known as Reflectance Transformation Imaging (RTI);

- The AHRC Noviodunum Archaeological Project, which has employed a number of formal archaeological data management tools, most prominently the Integrated Archaeological Database (IADB)
- The Százhalombatta Archaeological Expedition (SAX) project, which has employed alternative archaeological data recording methods including bespoke databases and Intrasis.

In addition to these and other specific projects our case studies reflect on-going research from within our postgraduate community, and in particular:

- Documentation of cultural heritage objects using three-dimensional surface and volume tools (laser scanning, photogrammetry, computed tomography);
- Workflow capture and representation (including case studies in ‘empirical provenance’ of Reflectance Transformation Imaging and blogging of computer graphic simulation development);
- The implications and potentials of Semantic Web approaches for archaeology;
- Contextual capture and delivery of archaeological data, including via mixed and augmented reality devices and wearable technologies;
- Digital capture and dissemination of informal and formal project meetings;
- Formal models of archaeological uncertainty and fuzzy data, including Allen operators for temporal reasoning and the potentials of the Open Provenance Model (OPM).

The IDMB project was able to bring together this broad expertise for the first time in order to create a coherent technological and policy approach. This represents a very considerable effort and is a major contribution of the project. It is our experience that such integration of experience across a discipline within our institution had not previously been undertaken. The approach taken could represent an effective template for aggregating specialist data management requirements and abilities in other disciplines. To date the archaeological case study has been introduced to research management across the Faculty of Humanities. The Faculty has also offered support to a roll out of the pilots and has funded additional evidence-gathering in terms of data-intensive Humanities practice across the institution. This will feed directly into university-wide practice.

What follows in this report is intended to represent best practice and requirements within the archaeology researchers in the institution. Whilst we inevitably draw on significant best practice, in particular from bodies such as the Archaeology Data Service (ADS), the present report focuses on the specific findings and innovations of IDMB.

## **Best practice findings**

The strategy adopted by the IDMB case study and the research projects that have informed it has taken into account a number of key issues. Where possible it draws upon the best contemporary archaeological data management approaches as well as applying recent advances pioneered in the UK and elsewhere. In the UK, our interest in mechanisms for documenting and managing archaeological data has drawn upon the outputs of ground-breaking research projects such as West Heslerton and the Silchester Project.

Archaeological fieldwork and post-excavation both generate a considerable range and quantity of data. In terms of size alone, the ADS ‘Big Data’ project has identified already that the issues of long-term preservation and access to the kinds of resources commonly created by methods such as

geophysics, computed tomography and digital photogrammetry are far from trivial. Moreover, many of these original datasets in turn are irrevocably transformed through a range of processes prior to completion. As a consequence, data management procedures are of central importance. Work such as the ADS's DAPPER Project provided a relatively early warning of the need for and issues implicit in preservation of such data.

One response to the clear need for large-scale digital resource management was an attempt to create single systems designed to manage all aspects of archaeological data. Some of the Southampton archaeological projects included in our sample have attempted to impose single data management infrastructures. However, there are some potential limitations of this strategy in the long term, particularly with regard to the scalability and flexibility of processes. Therefore in some other projects, notably the AHRC Portus Project, a range of interlinking solutions were employed, alongside to attempt to integrate them into a single, flexible workflow. Rather than developing a single, monolithic data management software tool, the Portus Project employed multiple solutions. This offered greater flexibility, with the accommodation of unforeseen needs and opportunities, and enabled continuing development and adoption of new technologies, whilst allowing users to interact with data in ways that suited them and the tasks in hand. It also provided project members with freedom to employ both proprietary and open formats, and both commercial and open-source software. The archaeological pilots for IDMB explored both approaches – the first a systematic, structured data management model (SharePoint 2010) and the second a more fluid model focussed on rapid implementation across diverse scenarios and expertise (EPrints).

Such flexibility to experiment is a crucial aspect of research-led archaeological computation. However, the lack of a single system governing computational practice on site and in the laboratory makes the implementation of formal data standards and processes crucial. IDMB demonstrated that standards are central to the way in which archaeological data are managed, ensuring that information as well as bits and bytes are retained, with data maintaining their meaning and relevance regardless of the contexts within which they are employed. Both in laying the foundations of a digital archive and in everyday practice, conformance to standards is essential. The use of metadata standards improves the accessibility of data in the short term and ensures that an archive resulting from this project will remain accessible and meaningful beyond the life of the excavation. The archive generated by the project must be sustainable, so that it can be interrogated long after the results are published. A number of key guidelines therefore have informed the IDMB archaeology data management pilots and policies. These include the ADS's guides to CAD, digital archives and geophysical data, and the English Heritage geophysical survey database.

Several archaeological content and asset management systems already exist that were examined during the IDMB project lifecycle. Our user studies suggested that these provide extremely well developed mechanisms for managing fieldwork information, but remain less well suited to the complex and heterogeneous asset archives that accompany and must interlink with them. The success of earlier exemplar systems such as G-Sys have prompted a succession of newer systems such as IADB, Intrasis and ARK. IADB is currently used by a number of UK-based archaeological field units, and on a number of archaeological research projects, including the University of Southampton's Noviodunum Archaeological Project in Romania, and Reading University's Silchester Insula IX Project in the UK). Intrasis, the Intra Site Information System has been employed widely in Europe and in the UK, including English Heritage at Richborough and Dover Castles. The ARK developed by L-P Archaeology is used by a range of archaeological projects and the latest release has received a great many downloads.

At Southampton we have captured archaeological data in each of these systems. Within IDMB we considered options for migrating the underlying data, evaluating the extent to which each application required maintenance of the platform in order to continue functionality. We also explored mechanisms for exposing the data held in these specialist repositories, including via web services and as linked data. The pilot implementations did not create links to these extant repositories but mechanisms were identified to create such linkages in the future. Above all the consideration of specialist content management systems demonstrated the limitations of monolithic data management structures and in part stimulated the rather sparse but effective three layer metadata model employed across the IDMB project. Whilst the ideal would be for all data to be seamlessly cross-referenceable whilst the technological means for this exist, and indeed their development is an area of research interest for the ACRG, the greatest challenge remains keeping track at a coarse level of data created in research practice and managing these resources for the long term. In the case of archaeology the latter issue is largely addressed by the exemplary record and expertise of the ADS. However, even for archaeology, prior to any formal deposit and where deposit of data with a disciplinary repository may be considered inappropriate (for example project management research data) the institutional repository remains key.

## Data type studies

Given the range of material described in our archaeology interviews and questionnaires we undertook a specific evaluation of best practice in terms of four core data types. These were digital photographic files, laser scanning data, RTI, and computer aided design (CAD) models. Whilst these are clearly not representative of the broad spread of data encountered in our surveys they provided a basis for assessing extant data management provision and for identifying areas for future research.

### Photographs

Archaeological research projects increasingly create enormous, unwieldy photographic archives. One single excavation project evaluated for IDMB included more than 30,000 photographs. Whilst this is in part a consequence of poor image management (limited readiness to delete) it also reflects the ubiquity of digital photography devices and increasing awareness of the potentials for comprehensive photographic coverage. For example, digital photogrammetry has enabled the metric reconstruction of archaeological sites where the extant survey data was insufficient. Photographs have increasingly become the preferred mode of navigating complex spatial data collections, and archaeology has been amongst the leaders in the adoption of such tools.

In terms of IDMB our core challenge was identifying mechanisms that could enhance extant practice and make better use of these photographic resources. In the first case descriptive metadata are assembled in a digital catalogue separate to the digital assets, frequently following an accepted metadata standard such as Dublin Core. This creates a portable, platform-independent archive that can be linked readily to other information. The second approach is to employ the metadata components built into the multimedia formats employed. For images these are commonly EXIF, XMP or IPTC data. Use of such embedded tagging ensures that the metadata are linked to the image, subject to accidental deletion via some image translation procedures and to limitations in access to the data imposed by the use of proprietary formats. Both metadata techniques were evidenced in the user survey and were implemented within the pilot, in both cases with the keyword metadata derived where possible from formal vocabularies. In some cases projects used both mechanisms. For example, capturing automatic metadata in the field via camera settings and geotagging, supplemented by manual (sometimes bulk) attribution of IPTC metadata in attribute:value pairs

using software such as Lightroom enabling offline editing, and finally supplemented by ingestion to an asset management facility such as MediaBin or the Sharepoint 2010 and EPrints pilots, all of which enable IPTC metadata to be accessed and augmented.

### **Laser scans**

Our institutional practice in terms of laser scanning data acquisition and documentation follows that defined by English Heritage. Some of our projects have also begun to consider the potential of different forms of annotation of the point cloud and surface data created, and also of the complex workflows involved in their processing. This parallels work by 3D-COFORM, MyExperiment and others. The pilot repositories were able to capture some of this detailed workflow information, with EPrints employing the standalone README.

### **RTI**

RTI data provide interesting data management challenges. As a department we have begun to capture large volumes of this imaging data and have undertaken some research on best mechanisms for managing and disseminating it. Within the remit of IDMB we considered the potential to draw broader conclusions from RTI for the management of large scale, multi-component data. In this we sought advice from Cultural Heritage Imaging and in particular explored what they have termed Empirical Provenance. The key points are:

- RTI datasets are the consequence of one or more related data capture events, that attract capture metadata; work on automated blogging of data capture devices is of relevance here;
- RTI datasets undergo a series of transformations; as with laser scan data above there is considerable potential for identifying formal workflows to document RTI data processing;
- RTI data are comprised of multiple files, with a need to identify connections between files and to audit transformations.

In the light of these requirements we generated sample RTI datasets for inclusion and evaluation in the pilot repositories.

### **CAD**

Archaeology along with disciplines such as Engineering and Architecture generates several forms of CAD data. Extant mechanisms for managing these at Southampton include bespoke CAD asset management tools such as Autodesk Vault, general media asset libraries with CAD plugins such as MediaBin, and generic repository systems. On-going work by CASPAR, CARARE and others is considering best practice in metadata attribution and workflow management for CAD resources. In the case of work by Southampton Archaeologists we have as part of IDMB considered the range of technological and policy methods that have been employed to document CAD data, and in particular their relationship to other forms of archaeological information. Whilst the pilots only implement relatively simple mechanisms for managing the connections between correlating sources, raw survey data and derived CAD models we have begun as part of IDMB to consider further options built around workflow management.

### **Sharepoint pilot**

The SharePoint 2010 IDMB pilot focussed on defining a set of classes suitable for recording highly structured data from across the institution, with an initial focus on archaeological fieldwork data. The aim of this was to provide a tailor-made solution matching the specific data structures of the

archaeological domain whilst requiring the limited bespoke programming. The archaeologists consulted wanted to continue to use their existing recording systems whilst linking to more comprehensive asset management tools that could be shared across the institution.

The archaeological implementation used the following data structure:

Super Class	Sub Class	Example Data and Values
Project		ProjectID:1 Title:Portus
	Project_Database (potentially inheriting from an unimplemented "Database" superclass)	ProjectID:1 Asset: Finances.xls Title: Financial summary for Portus budget codes Keywords: Finance
Sub-project		SubProjectID:1 ProjectID:1 Title: Imperial Palace 2007-2010
Area		AreaID:A SubProjectID:1
Trench		TrenchID:5 AreaID:A
	Trench_Photo	TrenchID:5 Asset: Keywords: Type: Record
Context		ContextID:1000 TrenchID:5
	Context_Photo	ContextID:1000 Asset: mycontextfile1.jpg Keywords: Coin Type: Record
Find		FindID:345 ContextID:1000 Title: Coin of Hadrianic date
	Find_Photo	FindID:345 Asset: myfile1.jpg Keywords: Coin Type: Snapshot
	Find_Photo	FindID:345 Asset:myfile2.jpg Keywords: Coin Type: Record

For the pilot implementation of this data model we used a subset of the AHRC Portus project dataset. The sample provided to the SharePoint development team included the following data types:

1. Context summary
2. Text documents e.g. Excavation report
3. Numeric data e.g. spreadsheet containing ceramic data

4. Exchange mail dataset e.g. research discussions
5. Photomosaic photos – geo-referenced
6. Balloon photos
7. Laser scan – time-of-flight data
8. Laser scan - video
9. Minolta 910 laser scan
10. CGI reconstruction model – original format
11. CGI reconstruction model – output animations
12. Building survey (CAD)
13. Plans (CAD)
14. Plans (scanned)
15. Sections (CAD)
16. Sections (scanned)
17. Surveyed elevations (CAD and scan)
18. Scanned paper architectural drawings
19. Photographs (including site photos, general photos, and publicity photos)
20. Aerial multispectral imagery
21. Aerial photography
22. GIS coverages (raster, vector and voxel)
23. Geophysical coverages (raster, vector and voxel)
24. CAD files
25. Finds photographs
26. RTI data of brick stamp
27. Drilled cores
28. Peter's model of Cistern complex
29. Drill-Coring data
30. Manual coring

SharePoint 2010 allowed the imported asset catalogue to be accessed through a range of views including Microsoft Pivot. It was able to implement the required hierarchical structure and provide visualisations of the underlying assets.

## **EPrints pilot**

The EPrints case study extended the archaeological fieldwork project needs explored in SharePoint 2010 in order to address the wider data management needs of the archaeology discipline at Southampton. Whilst the SharePoint 2010 pilot focussed on tight hierarchical structures mirroring and extending those employed by archaeological toolsets such as IADB, the EPrints case study was intended to offer a looser data management mechanism, supported by ingestion tools created by JISC DepositMO. The EPrints pilot data repository was based on a three-tier metadata system: project level, contextual level and data/ detail level. Within each tier the structure enabled one or more parallel hierarchical data structures to be implemented by implying structure from the order of keyword lists. This proved to be a very easy mechanism for capturing both hierarchically structured and largely unstructured data collections within archaeology.

The following screenshots provide an indication of the formatting for some sample geophysical data for the Portus Project.

## View Item: Geophysical data

This item is in review. It will not appear in the repository until it has been checked by an editor.

[Move to Repository](#)[Remove item \(with notification\)](#)[Return item \(with notification\)](#)[Preview](#)[Details](#)[Actions](#)[History](#)[Issues](#)

Kirstan, Stuart: Geophysical data: Raw geophysical data, grid locations, documentation <http://apropos.ecs.soton.ac.uk/166> (Unpublished)



**README file - Plain Text** (Format specification for geophysical grid file) [GeophysicalGridCoordinates.xlsx](#)

[Download \(8Kb\)](#)

This contains the format information for the geophysical grid data



**Data - Microsoft Excel (Excel 2010)** [GeophysicalGridCoordinates.xlsx](#)

[Download \(8Kb\)](#)

This file contains the grid data identifying the location of the individually 30m grid squares for magnetometry



**README file - Plain Text** (Format specification for geophysical data files) [DataReadme.txt](#)

[Download \(154Kb\)](#)

Contains detailed format information for Geoplot magnetometry data



**Data - Archive (ZIP)** (Zipped Geoplot Magnetometry Archive) [MagnetometryBundle.zip](#)

[Download \(187Kb\)](#)

Archive containing complete folder hierarchy for all magnetometry data gathered at Main Site - Area A. Note: zipped archive includes README file documenting folder hierarchy.

## Projects

[\[46\] Prinus Project Official URL](#)

**Item Type:** Dataset

**Subjects:** [C Auxiliary Sciences of History > CC Archaeology](#)

**Divisions:** [Faculty of Law, Arts and Social Sciences > School of Humanities](#)

**Depositing User:** Unnamed user with email [tdt@ecs.soton.ac.uk](mailto:tdt@ecs.soton.ac.uk)

**Last Modified:** 30 Aug 2011 08:27

**URI:** <http://apropos.ecs.soton.ac.uk/id/eprint/166>

**Data - Archive (ZIP)** (Zipped Geoplot Magnetometry Archive)  
 MagnetometryBundle.zip  
 187Kb

Hide options

---

**Content:**

?

---

**Format:**

?

---

**Format Description:**

?

---

**Context:**

1.
2.
3.
4.
5.

?

---

**Specific:**

1.
2.
3.
4.
5.

?

---

**Visible to:**

?

---

**License:**

?

---

**Embargo expiry date:**

Year:  Month:  Day:

?

---

**README:**

Archive containing complete folder hierarchy for all magnetometry data gathered at Main Site - Area A. Note: zipped archive includes README file documenting folder hierarchy.

?

## Archaeology pilot training for PhD students

### Survey Evidence

The results of our data management audit showed that many PhD students would find training and guidance on data management practice valuable. Students identified a need for support early in their research. Academics also acknowledged that this was an area where students needed skills training. Some academic staff noted a tendency for students to end up with dispersed data which were difficult to pull together, so endorsed early training to provide a foundation for good data management.<sup>1</sup> A link was also identified between training and continuity of support and practice. One archaeology student commented that there should be a “compulsory training course in the first semester with a small example project to work on as practice”. Another student said that “it would be nice to have a “toolbox” of Uni/help guidelines so you can pick what works for you”. As a result the approach taken for the training pilot was based around some core principles. Training should be:

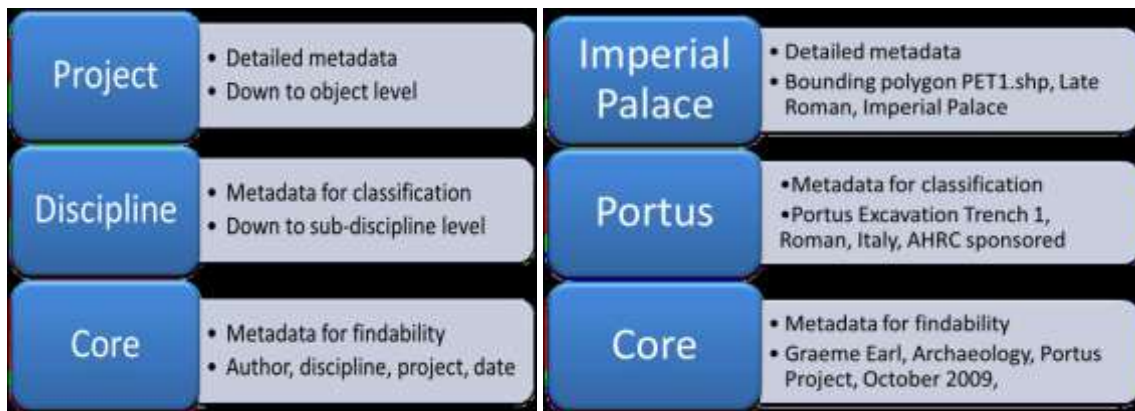
- Practice-led, linking with actual issues and examples from existing projects;
- collaborative, drawing on the expertise of students as well as academic and professional staff;

- integrated into the existing models of research training through research skills modules, incorporation into existing programme design and Graduate School provision
- extensible, so the model can roll out across academic programmes and services as part of a broader mix of point of need and embedded support.

## Practice-led

Training took the form of workshops and was issues-based. The approach aimed to work through a specific example from a project, looking at the storage/curation challenges and solutions and the roles of different stakeholders in managing the data. There was a strong emphasis on questions and discussion, encouraging peer sharing and learning. Final year PhD students and recent post-doctoral researchers could share recent relevant experience with those in their first semester of a masters or doctoral study.

Students were also introduced to the three-layer metadata model developed for the archaeology technical pilot and encouraged to think about how this would apply to their data as they started their fieldwork.



The JISC funded DMPTrain projects are now producing some useful training aids, case studies and resources, which were not available at the time of this pilot. The work to support archaeology from the DataTrain project is particularly relevant.<sup>ii</sup> We will be assessing these so we can maximise their usefulness to provide context and examples, whilst retaining a practice-led approach with current examples from activity within our research groups.

## Collaborative

The workshops were co-led by a member of academic staff and the liaison librarian for Archaeology. This built on the successful model used in the data management audit where interviews were jointly conducted by a post-doctoral researcher and the liaison librarian. The interviewers felt that this provided complementary expertise, with the researcher providing specific discipline knowledge and the librarian specialist expertise in metadata issues and policy. There was also evidence from the interviews that the actual involvement of a librarian in the interview had prompted the interviewee to reflect on the type of support that they might expect from the library, which included support for training.

## Integrated

The pilot incorporated the workshops into the MSc for Archaeological Computing as an example of specific embedding within a programme. They were also integrated into the research skills module which is taken by all archaeology masters students. A pilot workshop was developed for the Researcher Development and Graduate Centre (RDGC) aimed at PhD students and early career

researchers. The mid-term aim is to support the development of a full roll out of a data management training programme through the RDGC to support all disciplines. Initial feedback indicated that the approach taken in the pilot was too specific to archaeology and there were some issues with terminology and approach which hindered transference across disciplines. As a result of this we now plan to involve the University Strategic Research Groups (USRG) in the development of case studies. The USRGs are multidisciplinary groups aimed at investigating complex global challenges and we anticipate that this approach will both solve the problem of discipline specificity and provide cutting-edge examples that will engage students.

## Extensible impact

The overall approach has been to pilot the workshops in a way that acts as a template for implementation across the institution. Examples of data will differ, but the style of the workshops and the learning points can be extended to other disciplines. Feedback from the workshops will inform the next phase of development as we look to extend the range of programmes which embed data management training, particularly targeting relevant research skills modules. The USRG supported development of data management training through the RDGC will ensure that all PhD students and early career researchers are offered a core level of support. This extension of the pilot activity is a key part of the medium term implementation plan for the first phase of the Data Management Blueprint outlined by the project, where there is a commitment to “embedding data management training and support”. Other complementary mid-term goals are a “comprehensive and affordable backup service for all” and a whole “research data lifecycle” approach to data management.<sup>iii</sup>

## Next Steps

Through the data management audit and feedback from workshops students identified the need for point-of-need support as well as embedded training. We have earmarked this as a priority for the next phase. This will take the form of specialist support not suitable for training and build on the existing pockets of good practice identified by the AIDA assessment. It will include complex discipline specific queries and legal advice. It will also include tailored support for individual or project data management plans, storage options and metadata issues. To provide this support we will be enhancing our existing deskside support services<sup>iv</sup> and e-guidance. This will make use of existing tools such as the Digital Curation Centre resources to support data management plans.<sup>v</sup>

The success of our collaboration model of workshop delivery means that we are keen to build the capacity of professional staff to contribute to research data management training. As part of the next phase we would like to conduct an audit of training needs for staff, including librarians, research collaboration and bid support managers, and IT specialists. This will then inform the development of a tailored training programme.

These developments will continue to align with the overarching strategic approach. The recent restructuring within the University has consolidated the position of the RDGC as a central focus for co-ordination of training for the eight new Faculties. This presents a significant opportunity to extend this pilot work to the whole institution. We looked at cost-modelling specifically for training, but concluded that the pilot activity was too small a sample to scale up and cover the range of issues. We have, however, identified this as a key component of institutional cost and will look to include a detailed cost model for training as part of the next phase of business modelling.

## Summary of next phase

- Deskside training and one-stop-shop guidance
- Training needs analysis for professional staff
- Training programme available to all through the new RDGC
- Further embedding of data management training in MSc Programmes

- Development of case studies with the USRGs
- Full cost-modelling of data management training and support

Graeme Earl, Wendy White and Pam Wake, Aug 2011

The Institutional Data Management Blueprint project is funded by JISC under the Managing Research Data Programme. Acknowledgements: JISC programme manager, Simon Hodson; University of Southampton project contributors <http://www.southamptondata.org/about-us.html>

---

<sup>i</sup> Takeda, K. (2010) Initial Findings Report. <http://www.southamptondata.org/>, p.89.

<sup>ii</sup> DataTrain postgraduate teaching materials in managing research data, <http://archaeologydataservice.ac.uk/learning/DataTrain>

<sup>iii</sup> Takeda, K. (2010) Initial Findings Report. <http://www.southamptondata.org/>, p.4.

<sup>iv</sup> Deskside coaching booking, <http://www.soton.ac.uk/library/services/deskside/index.html>; <http://www.soton.ac.uk/isolutions/computing/training/deskside/index.html>

<sup>v</sup> DMP online, <http://www.dcc.ac.uk/resources/data-management-plans>