African Digital Scholarship & Curation 2009, Pretoria, 12-14 May 2009

A structured workflow for implementing digital archiving standards in an organisation

Peter MU Schmitz & Antony K Cooper

Logistics and Quantitative Methods, CSIR Built Environment
PO Box 395, Pretoria, 0001, South Africa
Email: {pschmitz, acooper}@csir.co.za

Topic: *Digital data management and curation*

Abstract

There are many standards available for archiving digital and analogue (paper) data and documents, but these come from various sources and it can be difficult to understand what standards should be used, and where. We have executed a project to improve the information standards in a government department, including those for data archiving. We have developed a structured workflow to guide the department on archiving their data and documents, which we report on here.

The workflow draws on the guidelines for archiving of electronic data, as developed by the National Archives and Records Services (NARS) of South Africa; various international standards (also adopted as South African National Standards); and the existing standards and guidelines in the department. Archiving activities in all government departments have to comply with the NARS guidelines. The guide starts with the generation of a file plan and ends with the backup or destruction of archived digital data (as appropriate) and the auditing thereof.

To facilitate using the workflow, we have prepared a detailed example for planning and conducting one of the department's key annual surveys. It shows where and when the various archiving steps should be taken. The survey plan consists of various milestones that need to be achieved when conducting the survey. The various archiving steps are then included in the appropriate milestones.

While the structured workflow was developed specifically for a government department, we feel that it is sufficiently general to have application for digital curation in other environments.

1. Introduction

There is a wide range of standards available to support the archiving of digital and analogue (paper) data and documents, but these come from various sources and it can be difficult to understand what standards should be used, and where. Within South Africa, the National Archives and Records Service (NARS) is responsible for preserving the "*national archival heritage for use by the government and people of South Africa*" and for "*promoting efficient, accountable and transparent government through the proper management and care of government records*" (NARS, 2008).

NARS promotes good archival and records management activities in general across the country and maintains an online index to several national and provincial archives, the National Automated Archival Information Retrieval System (NAAIRS – see http://www.national.archives.gov.za/naairs_content.htm). NARS also sets standards for the proper management and care of records still at national governmental bodies, which they term the *offices of origin* (NARS, 2008). These standards are of particular relevance for this paper, and cover:

- Designing and maintaining records classification systems;
- Identifying records with archival value;
- Determining the conditions for the management of micrographic and electronic records systems;
- Training public servants in records management; and
- Inspecting the records management practices of governmental offices (NARS, 2008).

Key documents from NARS include: *Performance criteria for records managers of governmental bodies* (NARS, 2003), *Records management policy manual* (NARS, 2004) and *Managing electronic records in governmental bodies – Policy, principles and requirements* (NARS, 2006). In addition to these, there many useful international standards from the International Organization for Standardization (ISO), many of which are also available as South African National Standards (SANS), covering detailed technical issues such as the permanence and durability of paper, binding documents, environmental conditions in archives, and electronic document formats. There are also guidelines and standards available from other countries and other international bodies.

All these standards and guidelines can be overwhelming and it can be difficult to understand how they all fit together.

The CSIR was awarded a contract by a government department to assist them complete several standards, including one for data archiving. Drawing on our experiences with implementing the Supply-Chain Operations Reference (SCOR) Model (Supply-Chain Council, 2008) for supply chain management in the production of geospatial data within a large organisation (Schmitz, 2008; Schmitz *et al,* 2007), as well as the discussion of workflow to structure and automate spatial data production activities as part of PhD research (Schmitz, 2008), we realised that it would probably be useful to use some form of structured workflow for this data archiving standard.

The department derives management information from a variety of sources, including electronic administrative systems and paper questionnaires. Hence, there are flows in various directions of both paper documents and electronic files (both over the Internet and recorded on discs), through several layers of the system.

The documents and files are also used by different sectors within the department. The documents and files are archived in various places and with different requirements for how long they should be retained. One complication is that the provinces function differently because of their different legislation and structures. For example, some of the archiving processes are centralised in some provinces but decentralised in others.

In the following sections we present an overview of the existing archiving standards relevant to our project, the structured workflow that we developed, and then we take an example through the workflow.

2. Existing archiving standards

Our review of the literature revealed examples of guideline documents on how to use the many archiving standards that are available and/or are imposed, for example, a publication on preservation strategies for digital libraries (Holdsworth, 2007) as well as the publications by the National Archives and Records Service (NARS) mentioned above (NARS, 2004; NARS, 2006), that aim to help governmental bodies in South Africa develop archives which conform to the Open Archival Information System (OAIS), published as ISO 14721:2003. The National Library of Australia developed PANDORA, which is a Web archive providing long-term access to online publications and web sites that have Australian content. The acquisition of archival material for PANDORA is via PANDAS (PANDORA Digital Archiving System) (NLA, 2008). PANDAS is a workflow-based archiving system that guides a user to archive material as well as providing a tracking capability for the user in the form of *Worktrays*, which gives the user an overview as to where the information is during the archival process (NLA, 2007). Hou *et al* (2006) have explored embedding automatically preservation processes into production video workflows, but those deal only with one type of medium (digital video) and do not cater for the human element in archiving.

With regards to archiving digital information in South Africa the following are important examples of archiving standards that are applicable for digital archiving in a government department:

- ISO 14721:2003, *Reference Model for an Open Archival Information System (OAIS)*.
- ISO/TR 15801:2005, *Electronic Imaging – Information stored electronically – Recommendations for trustworthiness and reliability* (published locally as SANS 15801:2005).
- ISO 18938:2008, *Imaging materials – Optical discs – Care and handling for extended storage*.
- ISO/TR 18492:2005, *Document Management Applications – Long term preservation of electronic document-based information*.
- ISO 19005-1:2005, *Document management – Electronic document file format for long term preservation. Part 1: Use of PDF 1.4 (PDF/A-1)* (published locally as SANS 19005-1:2006).
- ISO 15489-1:2001, *Information and documentation – Records management. Part 1: General* (published locally as SANS 15489-1:2004).
- ISO/TR 15489-2:2001, *Information and documentation – Records management. Part 2: Guidelines* (published locally as SANS 15489-2:2004).

Most of the above standards are embedded in the key NARS documents and the structured workflow diagram below (Figure 1) indicates which parts of the key NARS documents are applicable for each of the activities of the archiving process.

3. The structured workflow

Workflow is defined by the Workflow Management Coalition (WfMC, 1998, as quoted in Li and Coleman, 2004:4) as follows:

> *The automation of business processes during which documents and tasks are passed among participants according to a set of procedural rules and assigned roles.*

Business processes, according to the WfMC, are those linked sets of procedures and activities that are put in place to achieve the organisation's strategic objectives (Li and Coleman, 2004). Thus, workflow enables an organisation to capture information, processes and rules that are used to create documents and data with the aim of eliminating and/or reducing redundancies and time, and determining which part of the process can be automated (Koulopoulos, 1994).

Using the above descriptions with regards to workflow and the outcome of applying workflow and the SCOR model in a data environment (Schmitz, 2008), it was decided as mentioned in the introduction section to use the workflow approach to develop the archiving standard for the department.

Further searches on the Internet led to the Digital Curation Centre (DCC) Website. The DCC is an initiative funded by the Joint Information Systems Committee (JISC) in the United Kingdom (UK) that focuses its research on expertise in digital curation and on establishing best practice for the long-term storage, management and preservation of digital information with the aim that it is available for re-use over time. The DCC is a project based on the collaboration between the University of Edinburgh, the University of Glasgow, the University of Bath and the Council of the Central Laboratory of the Research Councils (Holdsworth, 2007). One of the main initiatives of the DCC is the establishment of a Digital Curation Manual to assist communities in preserving their digital information based on best practices known to the authors. This manual consists of various installments (chapters in the manual) such as *"Preservation Strategies for Digital Libraries"*, *Curation of Digital Art"*, *'Institutional Infrastructures for Digital Curation"* and *'Workflow"* (Digital Curation Centre, 2008a).

While 45 different installments have been identified, only 11 have been published to date. Unfortunately, the *Workflow* installment is currently only a shell on the DCC Website containing only the abstract and proposed key points. In their abstract, the DCC indicates the following:

> *In order to be successful, digital curation principles and approaches must be applied and undertaken throughout the complete lifetime of a digital object or item of digital information. The conception, storage, use and preservation of digital materials must all be undertaken according to the workflow plan that is best equipped to ensure longevity and sustained accessibility. An overview is required; a global understanding of the issues involved and of how and where these interrelate and influence one another (Digital Curation Centre, 2008b).*

The above reinforced that our approach to use workflow was the correct procedure to follow. Figure 1 below gives an overview of the workflow to archive management information in the department. As the standards are still drafts, we are not at liberty to provide specific details concerning the department, such as their existing standards relevant to archiving or the details of their surveys. We will not discuss the entire model here, only selected excerpts to illustrate the how the model integrates archiving processes into an existing departmental production plan for a survey.

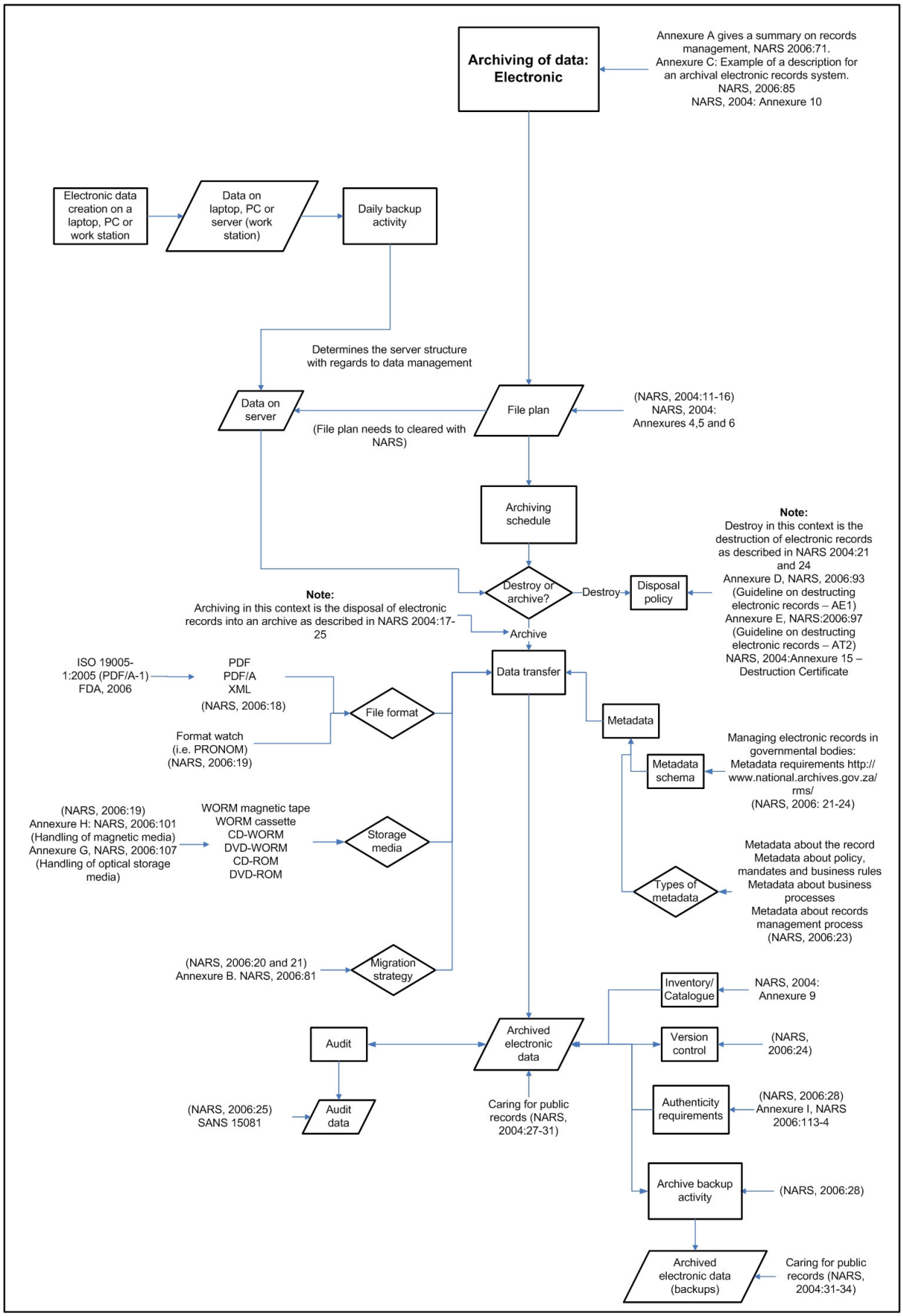

Figure 1: Process of archiving electronic records.

4. Detailed example

To facilitate using the workflow, we have prepared a detailed example for planning and conducting one of the department's key annual surveys. The example shows where and when the various archiving steps should be taken when the project is executed. The survey plan consists of various milestones that need to be achieved when conducting the survey. The various archiving steps are included in the appropriate milestones so as to guide the archiving process for this survey. For the purpose of the paper, only three excerpts from the example are presented here.

The survey is planned and executed based on various milestones that consist of tasks that need to be completed in order to achieve each milestone. The archival workflow was demonstrated using the survey plan. The survey uses hardcopy questionnaires to gather the required information, which are then captured electronically. Both the hardcopy questionnaires and the captured and subsequently processed information are archived. Only the archiving of the captured and processed data will be discussed here. The survey plan was provided to the authors in a spreadsheet. In the example prepared for the department, the various archiving steps were indicated in the milestone spreadsheet as well as in the flow diagrams that were developed to guide the archiving process. These flow diagrams are repeated below with the serial numbers of the various milestones or tasks in the tables at the appropriate places in the flow diagram. The underlined typeface indicates the milestones and the bold typeface in the various tables relate to archiving processes as indicated in the diagrams.

In this first milestone discussed, the file plans are determined in consultation with the relevant role players. These file plans are then used to archive the survey questionnaires (hardcopy) and the survey data (electronic). Figure 2 indicates the location in the workflow of tasks 1.5 and 1.6 from the extract of the survey plan shown in Table 1.

Table 1: Milestone 1.

	Start process	
No.	Milestones/tasks	Performance indicator
1	Information Needs Determination	Report compiled
1.1	Submit Inputs (Type A)	% Inputs Received
1.2	Submit Inputs (Type B)	% Inputs Received
1.3	Changes proposed after data analysis	Changes proposed are implemented
1.4	External Inputs	
1.5	**Setup file plan to archive hardcopy with registry**	**File plans should be determined in consultation with stakeholders**
1.6	**Setup file plan to archive electronic data with registry**	

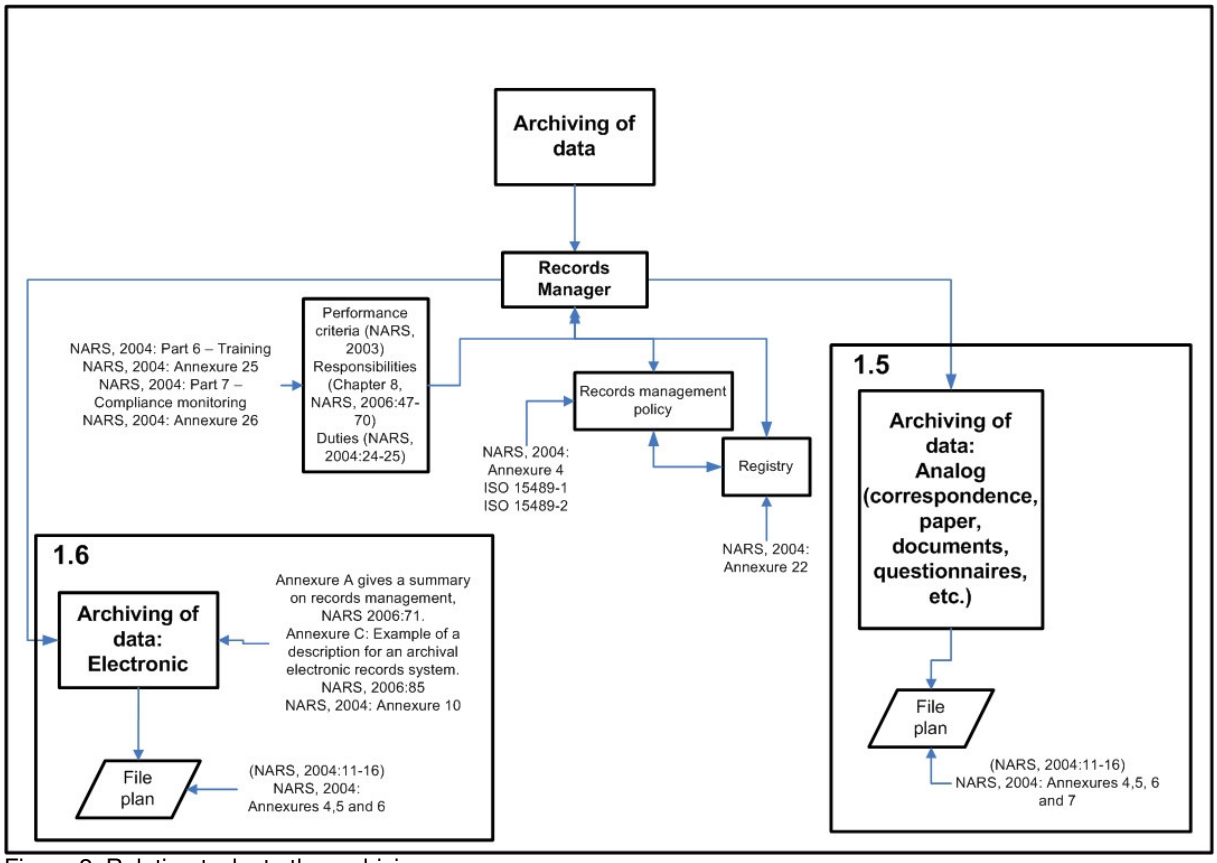

Figure 2: Relating tasks to the archiving process.

In the second example, Milestone 7 deals with the data capturing tool when it is being finalised and distributed to the units that capture online (see Table 2). Once the capturing tool has been finalised and distributed (Tasks 7.1 to 7.4), the department has to determine the file formats and storage media plus migration strategy to archive the captured questionnaires (Task 7.5).

Table 2: Milestone 7.

	Start process	
No.	Milestones/tasks	Performance indicator
7	Revision + Distribution of Electronic Capture Tool	Report compiled
7.1	Revisions made to electronic tool	
7.2	Testing of tool	
7.3	Documentation	
7.4	Distribution to units & user support	
	Guidelines for software usage	
	Training of Institutions on software tools	
7.5	**Determine file formats, and storage media for archiving including migration strategy**	

Figure 3 shows the location of Task 7.5 in the electronic data archiving flow diagram

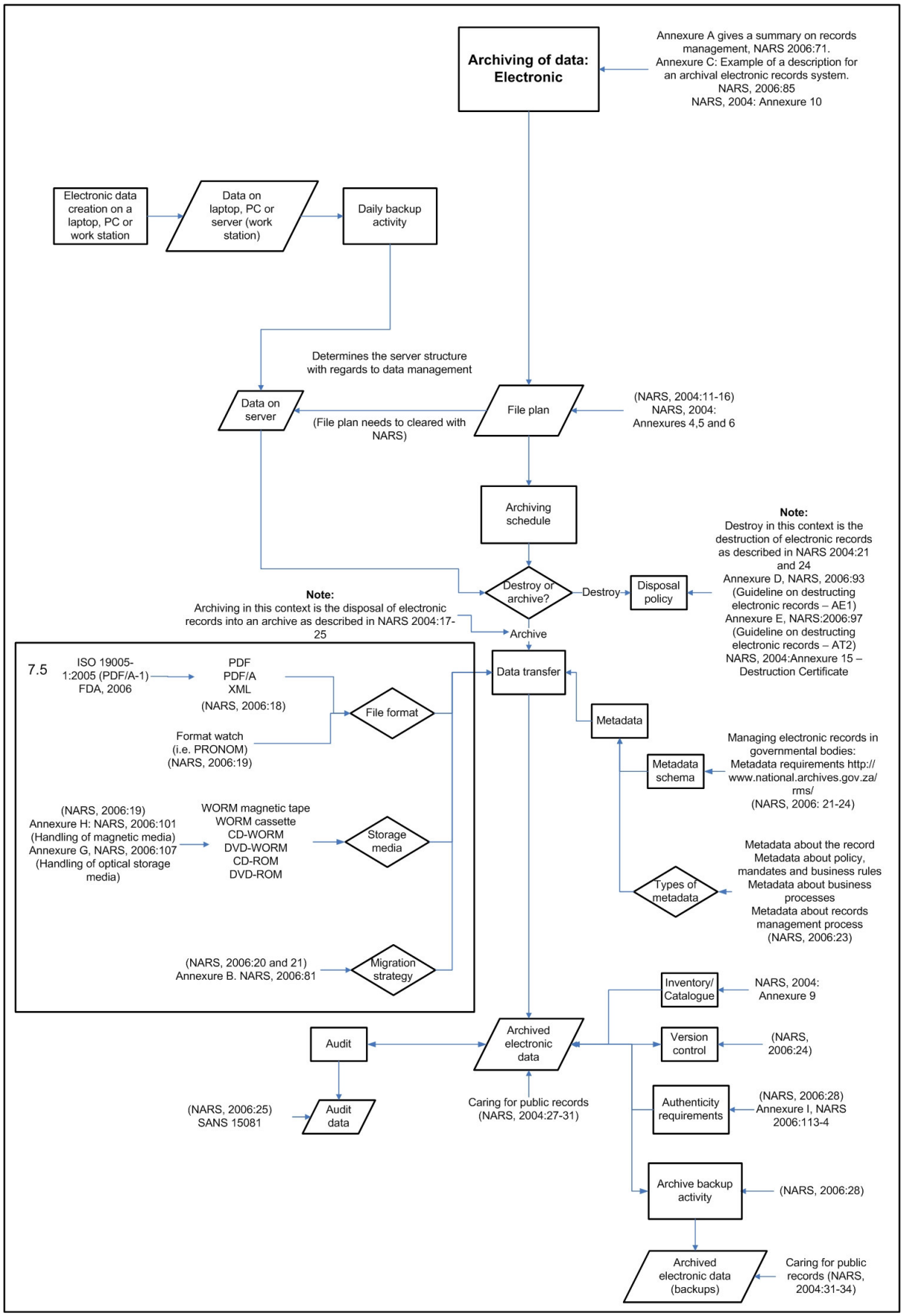

Figure 3: Task 7.5 with regards to the archiving of electronic data

The third example is Milestone 14, which is the last milestone of the survey process (see Table 3). Aside from the analyses of the data captured and the reporting, it is also the milestone where the electronic data and information about the survey (the metadata) are archived for future reference.

Task 14.5 concerns placing the data and information at the designated archive using the various file formats that have been determined during the execution of Task 7.5 (see Table 2 and Figure 3). Task 14.5 also includes the version control of the archived data. Task 14.6 concerns itself with the creation of the necessary metadata about the archived data and information. It is a requirement that the archived data and information are authenticated, so as to ensure that it is the definitive versions that have been archived, and this is done in Task 14.7. Task 14.8 deals with the backing up of the archived data so as to ensure that there is a copy available should the original archived be destroyed. Each archiving activity needs to be audited to ensure the integrity of the archiving of the survey data collected for a particular year (Task 14.9). Once the archiving process is complete, the archive inventory needs to be created or updated to reflect the contents of the electronic archive. These various tasks are shown in Figure 4.

Table 3: Milestone 14.

	Start process	
No.	Milestones/tasks	Performance indicator
14	Analysis & Reporting on the Data	Report compiled
14.1	Analysis of responses received as answers to Questions	
14.2	Documentation of the Above	
14.3	Compilation of Survey Change request	
14.4	Dissemination of data	
14.5	**Archive electronic data at designated archive**	
14.6	**Create metadata**	
14.7	**Authenticate archived data**	
14.8	**Determine archive backup strategy and create backups**	
14.9	**Archive audit**	
14.10	**Create/update archive inventory**	

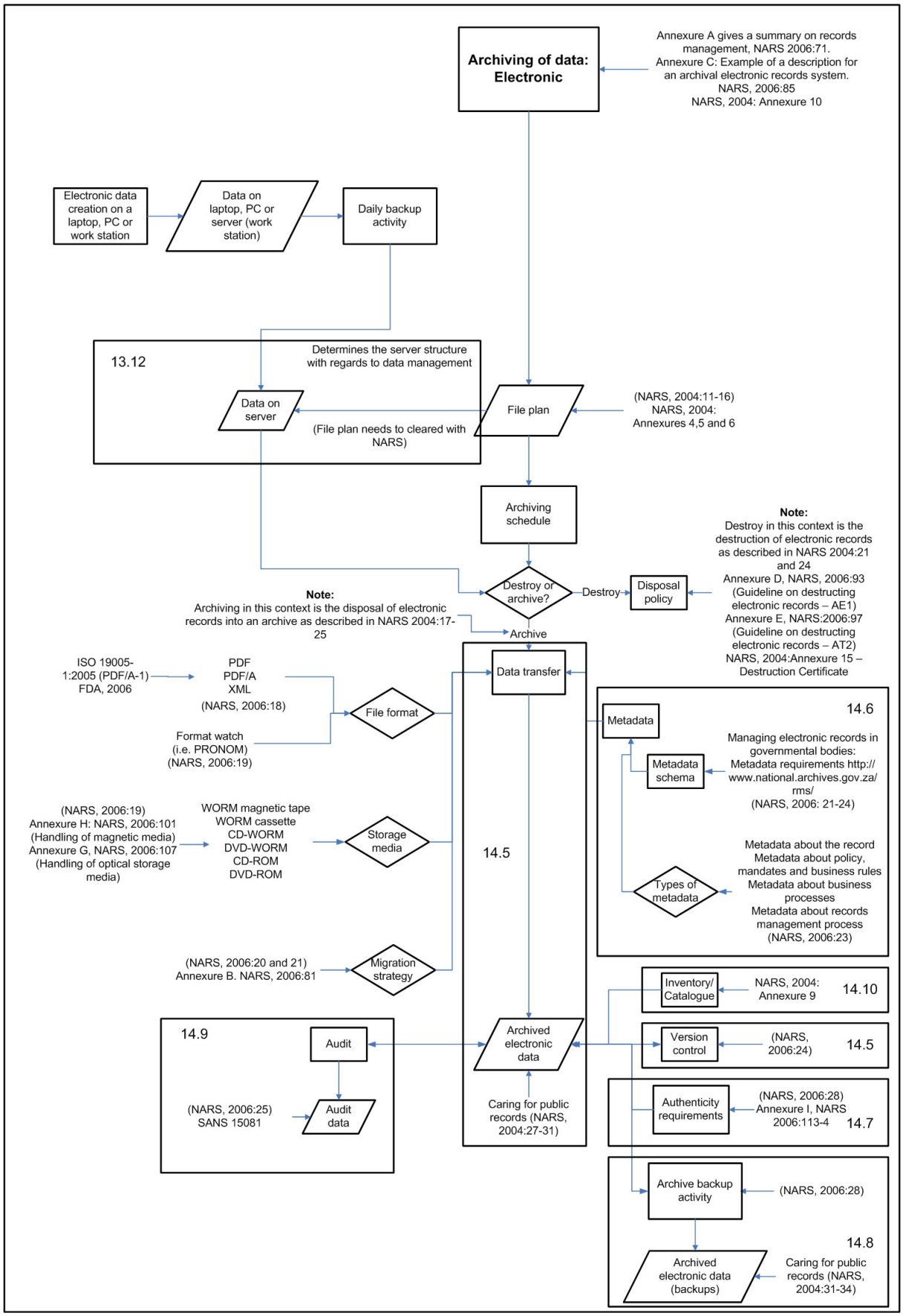

Figure 4: The archiving of electronic survey data and information

5. Conclusions

We have described here a structured workflow for implementing digital archiving standards in an organisation, which was developed as part of a project to improve the information standards in a government department. We feel that this workflow is sufficiently general to have application for digital curation in other environments. There are many standards available for archiving digital and analogue data and documents, but they come from various sources and it can be difficult to understand what standards should be used, and where. This workflow provides an easy to use outline of the processes to be followed when archiving digital data, indicating which standards need to be used for each of the process steps in the workflow. We have illustrated the workflow using experts from the plan for an annual survey.

6. Acknowledgements

The authors would like to thank the department for permission to make this presentation and we would like to thank the many staff members of the department for their contributions to the project. We would also like to acknowledge the contributions of our colleagues on the project at the CSIR.

7. References

Digital Curation Centre, 2008a: *DCC Digital Curation Manual Instalments*. Accessed 21 November 2008 at: http://www.dcc.ac.uk/resource/curation-manual/chapters

Digital Curation Centre, 2008b: *DCC Digital Curation Manual Instalments: Workflow.* Accessed 21 November 2008 at: http://www.dcc.ac.uk/resource/curation-manuals/chapters/workflow/

Holdsworth, D, 2008: *DCC Digital Curation Manual: Instalment on "Preservation Strategies for Digital Libraries".* DCC, ISSN 1747-1524. Accessed 21 November 2008 at: http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-strategies-digital-libaries

Hou, C.-Y.,Altintas, I., Jaeger-Frank, E., Gilbert, L., Moore, R., Rajasekar, A. & Marciano, R. (2006, June 20). A scientific workflow solution to the archiving of digital media. *The Workshop on Workflows in Support of Large-Scale Science*, held in conjunction with The 15th IEEE International Symposium on High Performance Distributed Computing (HPDC 2006), Paris, France. Accessed 14 November 2008 at: http://www.isi.edu/works06/hou-works06.pdf

ISO 14721:2003, *Reference Model for an Open Archival Information System (OAIS)*. International Organization for Standardization (ISO), Geneva, Switzerland.

Koulopoulos, T.M., 1994: *The Workflow Imperative: Using Workflow Technology to Create Adaptive Organisations and Enduring Competitive Advantage*. Delphi Publishing, Boston, USA.

Li, S. and Coleman, D., 2004: Modelling distributed GIS data production workflow. In *Computers, Environment and Urban Systems* (in press) Elsevier Ltd, p 25.

NARS. (2003). *Performance criteria for records managers of governmental bodies*. National Archives and Records Service of South Africa, Private Bag X236, Pretoria, 0001, South Africa.

NARS. (2004). *Records management policy manual*. National Archives and Records Service of South Africa, Private Bag X236, Pretoria, 0001, South Africa.

NARS. (2006). *Managing electronic records in governmental bodies – Policy, principles and requirements*. National Archives and Records Service of South Africa, Private Bag X236, Pretoria, 0001, South Africa.

NARS. (2008). *About the National Archives and Records Service of South Africa*. National Archives and Records Service of South Africa, Private Bag X236, Pretoria, 0001, South Africa. Accessed 14 November 2008 at: http://www.national.archives.gov.za/aboutnasa_content.html

NLA. (2007). *PANDAS WORKFLOW.* Accessed 21 November 2008 at: http://pandora.nla.gov.au/manual/pandas3/workflows.html

NLA. (2008). *PANDORA OVERVIEW.* Accessed 21 November 2008 at: http://pandora.nla.gov.au/overview.html

Schmitz, P.M.U. (2008). *The use of supply chains and supply chain management to improve the efficiency and effectiveness of GIS unit*, PhD thesis (unpublished), University of Johannesburg, 523 pp. Accessed 14 November 2008 at: http://researchspace.csir.co.za/dspace/handle/10204/2511

Schmitz, P.M.U., Scheepers, L., De Wit, P.W.C. & De la Rey, A. (2007, September). Understanding data supply chains by using the Supply-Chain Operations Reference (SCOR) model. *Logistics Research Network Annual Conference*, Hull, United Kingdom, 6pp. Accessed 14 November 2008 at: http://researchspace.csir.co.za/dspace/handle/10204/1441

Supply-Chain Council. (2008). *Supply-Chain Operations Reference Model: SCOR Overview Version 9.0*, Supply-Chain Council, Washington DC, USA. Accessed 14 November 2008 at: http://www.supply-chain.org/