



Delivery Notification for

Deliverable D12.1 SAM Website

Archive Storage and Technology

DOCUMENT IDENTIFIER	PS_WP12_BBC_D12.1_SAM_Website
DATE	07/03/2005
ABSTRACT	The Storage and Archive Management (SAM) part of PrestoSpace has produced a general introduction to storage technology, presented as a website : http://prestospace-sam.ssl.co.uk/ . The actual deliverable is the website. This submission is a covering document to formalise the fact that the website has been constructed and made available to the project, the public and the project reviewers.
KEYWORDS	website storage technology archive
WORKPACKAGE / TASK	WP12.1
AUTHOR, COMPANY	Richard Wright, BBC
NATURE	Report (actual report is the web site).
DISSEMINATION	Public
INTERNAL REVIEWERS	David Jordan, BBC; INA

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
1.0	07/03/05	First version of delivery notification	Living	Confidential
1.1	09/03/05	Final review of delivery notification	Notified	Public

1. Document Scope

The Storage and Archive Management (SAM) part of PrestoSpace has produced a general introduction to storage technology, presented as a website. The actual deliverable is the website, and this document is just a covering document to formalise the fact that the website has been constructed and made available to the project, the public and the project reviewers.

2. Summary of Deliverable

The intent is a Web site for the partners and the public alike, which will be the first point-of-call for information on audiovisual storage. The key issue is sustainability: keeping the information up-to-date. The project will continue to work with the industry over the entire course of the project, to keep the information up-to-date. The intent is to make the site 'the place to go' for archive storage high-level information, so that the site will continue beyond the duration of the project.

The website supplies information and management tools on digital technology for the storage of film, video and audio content and associated metadata. The information covers the state-of-the-art in storage technology, and includes forecasts of trends over the next ten years. The information has been gained from the requirements and experience of the consortium partners, open literature, from a survey, and from direct contact with leading technology companies in the field.

At present the website is hosted by the technical developers SSL (a PrestoSpace and SAM partner). It is fully functioning, but the website will (by the completion of the project, and probably sooner) be incorporated into the general PrestoSpace website.

<http://prestospace-sam.ssl.co.uk/>

There is a detailed plan, in Workpackages 12, 13 and 14, for the further development of this website during the course of PrestoSpace. We are also working on various possibilities for the long-term sustainability of the website (through commercial sponsorship or by being 'adopted' by a host institution).

3. Notification of Delivery

The URL for the first version of the website was made available to the PrestoSpace project in early February 2005. It has been reviewed at a SAM meeting (02.03.2005) and at a meeting of the storage industry (01.03.2005). Initial responses have been favourable.

Problems: The site includes description of future components of the site: future deliverables. This 'advertising of what we haven't done yet' may cause user frustration. We will monitor feedback and adjust the site as needed.

4. Detailed Description of Deliverable

The following sections describe the content of the deliverable. The technology used to produce the website could have an equally detailed description. The website uses some sophisticated tools to make it more than simple pages:

- All pages are generated by a content management system. This allows the 'side bars' of the pages to change in tandem with the central portion, always providing relevant material.
- All content is tagged with keywords, to allow the above-mentioned 'associative' functionality, and also to improve search results.
- There is a thesaurus to control and define the keywords, and assist navigation and search functions.
- The site uses RSS feeds to constantly capture new, relevant material from other websites, and add them to the knowledge base.
- The content we have developed plus the captured content are held in a knowledge base, searchable by thesaurus / keyword as well as free text.
- The site incorporates calculation functionality, to implement the user-support calculation aids ('ready reckoners'): deliverables 12.2, 13.2 and 14.2,

4.1. Description of Content:

I. Summary:

We have a workpackage-based plan for three sets of information and three calculating tools. These are briefly described, but possibly of greater interest (and complexity) are their interactions. The concluding section describes possibilities for having separate formal deliverables, forming an integrated website.

II. Goals of the workpackages:

There are three SAM workpackages, each with an informative deliverable (a website and two reports, and each informative deliverable has an associated calculating tool. The workpackages cover the following three areas:

1. **Basic storage information.** The deliverable is the website, and within the website (if possible) is a calculating tool about volumes and costs.
 - A branching set of option, organised as a tree or walk
 - Business-purpose related information along the branches (along the walk) relating to the decisions **between** options; comparative information between two alternatives.
 - Detailed information at the end of the paths (leaves of the tree), covering everything we can reasonably assemble about that 'endgame' – descriptive information about one area.
 - Tables or lists of information about the commercial sector:
 - Storage information
2. **Business case information** – for a preservation project. We have a detailed document from ITI as the information deliverable, and again on the website a calculating tool relevant to a

business case. The business case calculation will involve (among much else; pruning to make a realistic-but-still-simple calculator will be essential).

- Volumes (of material)
- Process – the workflow applied to the material
- Outputs – what is produced by the process
- Storage – a place to put the outputs, using the option (the leaf or endgame) determined in part 1. Here is the first complexity: the walk in part 1 doesn't necessitate any user input of volumes, conditions, formats. Hence part 1 could be a coherent activity – a decision-making activity – without involving anything about storage calculations, and without involving any user data. But: one vital part of decision-making is comparative cost.
- TCO – Total Cost of Ownership which we'd like to estimate for 20 years. TCO is really only relevant to the hosting of the 'outputs'. The initial volumes disappear (though we could include an optional cost for 'keeping the originals', the process is a revenue cost (except for equipment, but generally the equipment would be written off during the project) and the outputs may have a one-off media cost but replenishing media (migration) and any commensurate life-cycle issues in other technology should all come under Storage.
- ROI – Return On Investment information has been a requirement of the PrestoSpace project since the very early days of the EC Sixth Framework, where ROI considerations for cultural heritage investment was a general target of the research programme. The essence of the difference between Workpackages 12 and 13 is in my view ROI.

Everything about a user's (a website user's) storage costs could be included in part 1. The alternative is to keep part 1 general, maybe dealing just with a few archetype cases rather than getting the user's details – if we need be more specific. If so, all user detail is entered as part of the business-case building exercise.

3. **Project Logistics** – support in running a preservation project. The BBC deliverable 14.1 (in preparation) outlines this Work Area. There is a wealth of information in the experience of the BBC and other PrestoSpace partners, but this has not been compiled. The essential planning and tracking information includes:
 - Budget
 - Staff
 - Facilities: equipment and workspace
 - Storage
 - Transport
 - Volumes: material in and out

III. Fulfilling the goals: preliminary results

1. **Basic storage information.**

We have a basic decision tree, as follows:

- A Media on shelves: Optical or Magnetic
- 1 Optical: AV or IT format
 - AV formats: Audio CD or video DVD
 - IT formats: CD-ROM, DVD-ROM
 - 2 Magnetic: Disc or Tape (Disc on shelves?? Really??)
 - Disc: various kinds of discs and disc interfaces
 - Tape: Videotape, Datatape,
 - Videotape:
 - Conventional formats
 - IMX "file on videotape"
 - Datatape: linear or helical
 - Type of computer interface:

At the end of the decision tree, the information provided should centre on total cost of ownership, including cost to write, storage, migration cost, time interval between migrations and any other maintenance costs – all additional factors beyond raw media cost.

- B Mass Storage
- 1 In-house
 - Tape Robots or Disc Arrays (but these can be used together, so not an either-or)
 - Types of tape robot (cf types of tape !)
 - Types of Disc interface:
 - Direct: IDE, SATA
 - Special interface: SCSI, Fibre-Channel, USB 2
 - Networked: SAN, NAS, iSCSI, wireless (??)
 - 2 External
 - 2A Within the Enterprise
 - 2B Outsourced
 - 2Bi Commercial
 - 2Bii Public Service

In addition, we should provide what we can of the following basic or background material, as general information rather than keyed to position in a walkthrough (though it could be cross-referenced to other areas of the site)

- C. Tables or lists of information about the commercial sector:
- Manufacturers
 - Suppliers
 - System integrators
 - Consultants
 - Suppliers of outsourced storage
- D. Storage information
- Media
 - Media integration – robots, RAID, MAID ,,,
 - Media management – integrity checking, media backup and migration
 - Connectivity – interfacing and networking
 - Access

2. Business case information

In Part II Goals, we listed the business case elements as:

- Volumes (of material)
- Process – the workflow applied to the material
- Outputs – what is produced by the process
- Storage TCO – Total Cost of Ownership.
- ROI – Return On Investment

We have a choice of getting ‘volumes’ information in for ‘Basic Storage Information’ purposes, or somehow walking through the storage options without that information, and collecting it here. The remainder of the business case information could be collected here, as part of a walk-through of the constituents of a business case, as outlined by ITI in D13.1.

Volumes, Process, Outputs and “Storage Choice” information are inputs to the business case building and calculating activity; TCO and ROI are calculated outputs – though the most useful output will likely be forcing the user to get all the inputs ‘present and accounted for’. If volumes data has already been collected for determining “Storage Choice”, then it will be available (if we can keep it; the user may be in a new ‘session’ when returning to look at business cases) and the new inputs needed are the Process parameters and Output choices.

3. Project Logistics

As discussed at the September SAM meeting, a process tracker could look like a system simulation, with various elements linked to perform a function (media preservation), and consuming resources (and making new assets) during the performing of that function (process).

As stated above, the parameters / dimensions are:

- Budget
- Staff
- Facilities: equipment and workspace
- Storage
- Transport
- Volumes: material in and out

For tracking, it is likely that a user will want to have periodical data – movements, costs, outputs every week, month or year.

There is a general issue of the planning of these dimensions, and then the tracking of performance vs plan. This requires updated forecast data with actual data, which probably more than doubles the complexity of an already difficult situation with regard to obtaining, retaining and returning user data.

IV. Calculation requirements

1. STORAGE: Comparative costs supporting decisions in part 1
 - i. Based on a few archetype collections (1k, 10k, 100k, 1M items); level and kinds of access could be archetype parameters, or could be just parts of dialogue at decision points)
 - ii. Based on user-supplied numbers
2. CASE: Building a business case
 - If no user information from STORAGE, then
 - Input: Volumes, Process, Outputs and “Storage Choice”
 - Output: all the above tidied up; plus TCO; plus ROI
 - If STORAGE has Volumes data, then input only of Process, Outputs

Output: all the above tidied up; plus TCO; plus ROI

3. TRACK: Project planning and detailed tracking, of dimension covering:
 - Budget
 - Staff
 - Facilities: equipment and workspace
 - Storage
 - Transport
 - Volumes

V. An integrated approach

Although the process tracker could look very simple, and appears to come at the end of a chain of thought, data gathering and calculation that leads through all the above, it could be made the 'way in' to the whole scenario.

There could be a 'demon' who changes old media into new, and in order to get that demon to represent something closer to a particular situation (ie the real-world situation of a user of the website) the user will have to (will be able to!) change the parameters of the demon, to get more and more relevant results.

In the simplest case, a user comes to a page where a demon is presented as being able to carry out a preservation project. The demon has a stack of old media and a Start button. There is a clock and a cash register. Press Start, and the demon transfers old media to new, each time ringing a charge on the cash register and registering additional time on the clock. The process stops when the old material is consumed.



For the user to get some more relevant information, the user clicks on one of the icons to change the parameters of that icon:

- Volumes information is at the input stack: how many, of what.
- Cost information is at the cash register: media and staff costs; options for facility and transport and IT costs.

Process information is at the demon (the little man): what the demon is doing, and the outputs. It is here that the demon can be broken down into sub-demons covering the elements of the workflow, including getting information from the catalogue, finding the old media, transport, labelling, actual transfer, transport again, updating the catalogue, quality control (one or more

stages of quality control). The demon can also be replicated to add multiple streams of production, and the demon can have simple parameters for dealing with problem material.

- There will HAVE to be a “time per transfer” parameter value for the demon, and we should suggest that material that causes problems goes to a B-stream demon (B-stream sub process) which has a separate cost per item and throughput, and failure rate.
- Output information is also part of the demon – what the demon makes: formats and multiple formats, how big and how costly.
- Storage option is also a demon parameter. Possibly Storage option and outputs collapse into one.
- Finally, there are control parameters for running until time or materials are exhausted, or for readout of process parameters at time intervals: by clicking on the Start Button icon.

SAM will explore the possibility of achieving all three sets of workpackage goals – at least in so far as they relate to web resources – as details reached via the demon. Certainly all the WP14 goals should be straightforward, but it may be that *business case* (to get a budget for the cash register) and *storage options* (to identify that part of the demon’s output, and get cost and volumes data) can also meaningfully be linked.

This is not to say there shouldn’t also be separate walk-throughs. The primary advantage of the demon is that it creates a page where everything comes together.

5. Related Documents and Links

The current (March 2005) URL for the SAM website is:

<http://prestospace-sam.ssl.co.uk/>