PLANNING AND MANAGING THE 'COST OF COMPROMISE' FOR AV RETENTION AND ACCESS

M.J. Addis, M. Jacyno, M. Hall-May, M. McArdle, S.C. Phillips

University of Southampton IT Innovation Centre, UK

ABSTRACT

Long term retention and access to AV assets as part of a preservation strategy inevitably involves some form of compromise in order to achieve acceptable levels of cost, throughput, quality and many other parameters. Examples include: quality control and throughput in media transfer chains; data safety and accessibility in digital storage systems; and service levels for ingest and access for archive functions delivered as services. We present new software tools and frameworks developed in the PrestoPRIME project that allow these compromises to be quantitatively assessed, planned and managed for file-based AV assets. Our focus is how to give an archive an assurance that when they design and operate a preservation strategy as a set of services that it will function as expected and can cope with the inevitable and often unpredictable variations that will happen in operation. This includes being able to do cost projections, sensitivity analysis, simulation of 'disaster scenarios', and to govern preservation services using Service Level Agreements and policies.

INTRODUCTION AND MOTIVATION

This paper presents an approach and supporting software tools/framework for the planning and management of audiovisual preservation of file-based AV assets. Long term retention and access to AV assets inevitably involves compromise in order to achieve acceptable costs, especially for the huge volumes of archive material often involved. But how can these compromises be objectively and quantitatively assessed? How can an archive be assured that when they design and operate a preservation and access infrastructure that it will function as expected and can cope with the inevitable yet unpredictable variations that will happen in operation? For example, how can the infrastructure be provisioned or managed so it is robust or flexible enough to cope with variations that happen in everyday operation (e.g. volume of AV material to be handled, availability of operators and other resources, demand for archive access). How can 'disaster scenarios' (e.g. large scale storage failures, step-changes in workload, sudden loss of staff) be simulated and planned for? How can the functional services of the infrastructure (ingest, access, storage, replication, fixity checking, etc.) be monitored and managed using defined SLAs for the different users of the system whilst ensuring internal resources are maintained for essential preservation actions (e.g. migration, fixity checking, metadata and format validation during ingest).

Effective planning and management of a preservation infrastructure is increasingly important as audiovisual archives become file-based and an active element of the production, post-production and distribution process. Often archive systems are in-house, but increasingly parts are out-sourced and even off-site. We use policy-based planning and automation applied to outwardly facing archive services and internal preservation processes alike, defined through SLAs and actively measured and controlled against

metrics for performance, data integrity and availability.

APPROACH

Our approach to planning and managing services for preservation and access is shown in Figure 1.

The application channel at the bottom contains the services that deliver preservation and access, e.g. the tools and services that would be found within the main functional areas of OAIS [1]. By considering the application channel as a set of services, each of which has an SLA and defined Quality of Service (QoS), then each service can be monitored and managed consistently.

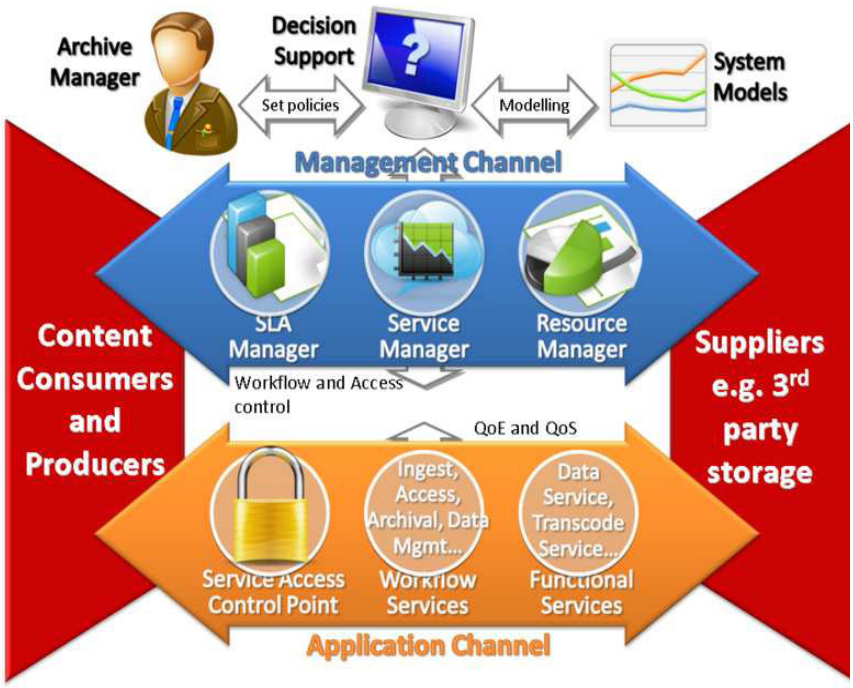

Figure 1 Planning and Management approach

The management channel automates the management of these services and includes customer and supplier relationships. This applies equally to in-house deployments, e.g. within a single organisation, or when using third-party providers. Within the management channel, the SLA manager deals with customer SLAs (those of the consumer and producer in OAIS terminology) which set out the constraints and service level objectives (SLOs) on ingest and access. The Resource manager deals with supplier SLAs (such as out-sourced storage or compute facilities) and with in-house resources. The Service manager balances commitments to customers with the resources available internally and from external suppliers. An event-decision-action loop makes decisions according to a set of policies.

The decision support tools at the top of the figure are where people design, test and set the policies to be executed by the automatic management layer. This includes planning and simulation of 'what if' scenarios. The output of the decision support tools are costs, plans and a set of management policies. The service manager in the management channel then uses the policies to decide which actions to take in response to observed behaviour of the application channel services, for example bottlenecks, failures and a lack of resources.

The key to our approach is the loose coupling of the layers using a very simple and light-weight interface. This is crucial for deployment in a wide range of practical settings. Example scenarios we target using our tools and framework include:

1. Distributed file storage systems that need proactive management to ensure an optimum balance of data safety, accessibility and cost.
2. AV migration, e.g. file format migrations or transfer from discrete media to digital files. A balance often has to be struck between quality, throughput and cost.
3. Ingest and access using performance KPIs. Here issues are cost, performance, user prioritisation, impact on other archive activities e.g. ingest and maintenance.

In each case, the scenarios involve some element of unpredictability, e.g. because a process is stochastic or because the real world workload on a service will be very variable and hard to predict. This means whatever initial plan there might be for a given scenario will need active monitoring and management at 'run time'. In each case, the scenario involves some form of trade-off (e.g. between cost, quality and throughput for a digitisation chain) and there is the need for optimisation, both at the planning stage (e.g. how many QC operators to use, whether to use software-based video defect detection) and at the operational management stage (e.g. load balancing, addition of more QC stations etc.).

SIMULATION AND MODELLING

Cost, loss and resource planning in an archive storage system

Much work has already been done on the cost and reliability of storage systems, including for preservation of audiovisual content [2]. Google [3], San Diego Super Computing Centre [4] and others report the Total Cost of Ownership (TCO) including analysis of how this falls over time. There are many reports on the reliability (or lack of it) for storage technology and storage systems, including the types and origins of failures [5], mostly for Hard Disk Drive (HDD) based systems, but also field studies and evidence of failure rates seen in practice [6] including for AV archives [7]. However, there is relatively little work that investigates the trade-offs that exist between cost, loss and ease of access [8]. There are many choices that can be made, for example number of copies to make, what technology to store them on, how often to check their integrity, whether to use automated or manual processes, and how to balance user needs (e.g. ingest and access) with internal functions (e.g. media migrations, file scrubbing, replication). These activities take time to execute dependent on resources available (e.g. people, servers, bandwidth), which in turn cost money and become contended with different uses having different priorities.

Our approach to simulating this problem starts with a simple but flexible storage model (Figure 2) that has the function of accepting files for storage (writes), returning files from storage (reads) and storing the data inside the files using some form of physical media (hard drive, data tape, optical disc etc.). The model includes a 'controller' (manual or automatic) that mediates these processes.

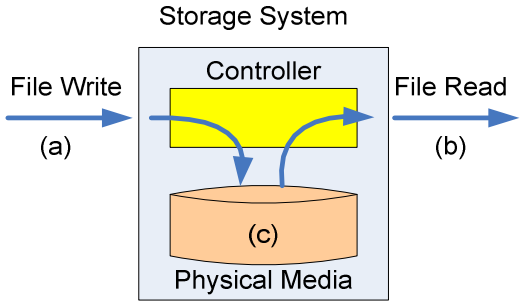

Figure 2 Storage model

The model can be applied to automated hardware/software, e.g. a HDD server, or it can be applied to a more manual process, e.g. data tapes on shelves with archive staff that put new tapes onto shelves and retrieve existing tapes to serve user access requests. When writing or reading files, various operations may be applied, e.g. encoding or applying error correction. Depending on the system being modelled, this could be by firmware on a HDD, the RAID controller in a HDD array, integrity management in a ZFS filesystem, manual integrity verification by an operator, or a combination of all of these. Likewise, various failures or errors could occur, both latent or extant, which could range from 'bit rot' in a HDD system through to accidental damage from manual handling of discrete media. These can happen (a) when data is written, (b) when data is read, and (c) when the data on the physical media is in effect 'doing nothing'. These are all represented through error rates for read/write/store actions. The actions each has a cost, which forms the basis of the associated cost model (one off ingest cost per file when adding it to a storage system, access cost per file incurred each and every time it is retrieved from the storage system, and storage cost per file when it is inside the storage system with the cost being a function of how long the file

has been stored for).

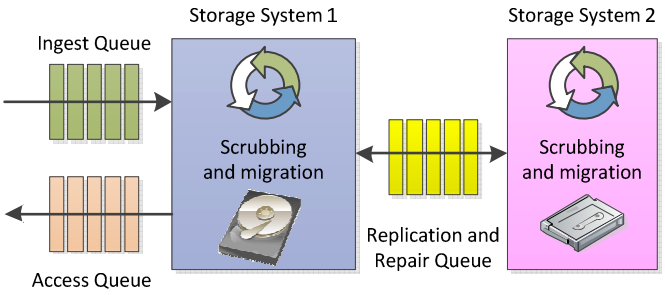

Figure 3 Archive storage configuration

One or more storage systems are then combined into an archive configuration. An example is shown in Figure 3. This includes ingest and access queues for new file arrival and retrieval of stored files. The configuration determines how files are allocated to storage, how they are replicated, and how they are repaired if there are failures. Resources can be allocated to serving ingest, access and copy operations as well as for activities within each storage system, e.g. integrity checking and repair. A set of template configurations are provided that correspond to common patterns for real world storage configurations, e.g. mirrored servers, HSM, online + deep archive.

The interactive simulation tool takes a discrete event simulation approach. During the simulation, time ticks away (e.g. 1 second of the simulation might correspond to 1 week in the real world) and events are generated (e.g. random corruption of files in a storage system, requests to access a file, new files to be added to the archive). These events then trigger actions, e.g. a copy/repair process, which is then added to the queues of the storage systems involved. A storage system processes items in its queues according to how much resource it has available (e.g. serving access requests sequentially or in parallel). The available capacity of the resources used by each service determines how many items are processed for each tick of the clock, and at what cost.

The user can interact with the simulation as it progresses, e.g. changing the amount of resources available or changing the policy for data safety (e.g. making more copies or checking them more often). In this way, the user is in effect playing a game that helps them understand how to react to and manage events that they might see in practice when operating a real system. For example, there is also an option to simulate 'disaster scenarios': rare but catastrophic events where large fractions of the storage become temporarily or permanently unavailable. An example of a simulation is shown in Figure 4 for a 2 copy model with periodic scrubbing. This includes simulation of an unexpected corruption event (1% of files lost at start of Sep 2010) that causes an overload on resources with consequent file loss and considerable time before the overall system returns to a stable operating state of little or no further file loss.

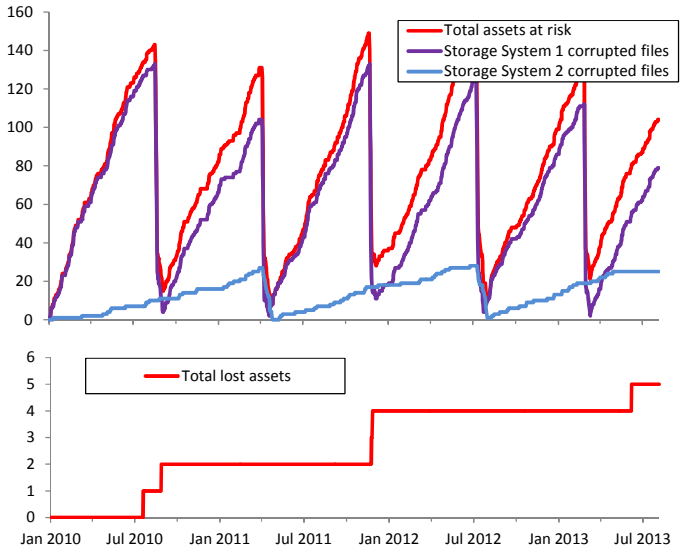

Figure 4 Storage and access simulation

Cost of quality and throughput in a D3 transfer and QC workflow

The D3 project at the BBC is an effort to migrate video from approx. 100,000 D3 tapes into file format (MXF wrapped uncompressed video and audio) and store it on LTO data tape. Technical details can be found in the BBC whitepaper 155 [9] and is shown in Figure 5.

The workflow starts with D3 tapes that operators load in to D3 decks and capture the resulting SDI stream to a file. These files are then written to data tape and the AV content manually inspected by QC operators at dedicated QC stations. The operators look for defects introduced during the transfer as well as already existing in the video (e.g. from previous migrations such as 2" Quad to D3).

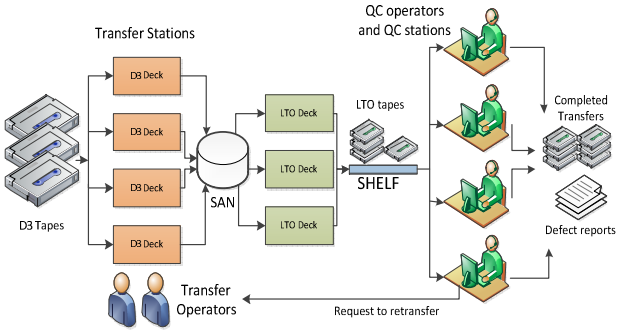

Figure 5 D3 transfer and QC workflow

Inputs to the simulation include the number of D3 tapes, the number and cost of D3 operators and decks, the number and cost of QC operators and workstations, the time and resources needed for each step (e.g. transfer, reviewing defects), the frequency of defects and the effectiveness of operators in detecting them, the likelihood of retransfers being required, and the cost/capacity of the storage systems used in the workflow.

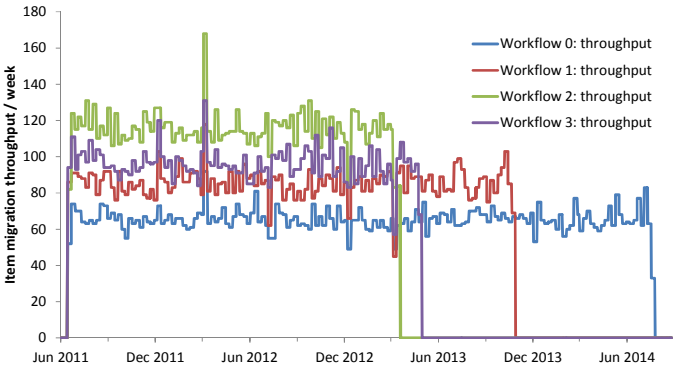

Figure 6 Workflow throughput analysis

Scenarios that can be simulated include: (a) the result of reducing time spent on manual QC, e.g. time-boxing instead of a full pass of every item (b) the benefits of using automated quality analysis software to guide the QC operators, and (c) the effect of increasing resource to remove bottlenecks, or the impact of temporary loss of resources, e.g. operator illness or systems failures.

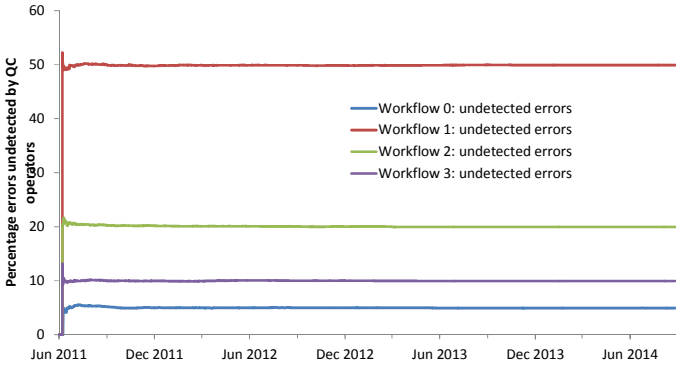

Figure 7 QC efficiency analysis

The result of a typical analysis is shown in Figure 6 which shows the rate at which D3 tapes complete the process for different workflow configurations and Figure 7, which shows the corresponding number of defects not picked up in QC. The costs of the different configurations can be compared and hence a cost/ throughput/ quality comparison done.

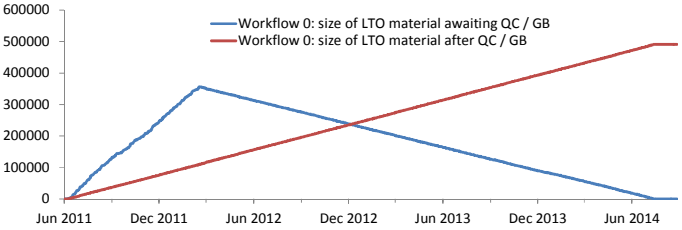

Figure 8 Resourcing analysis for the QC step

Optimisation and sensitivity analysis can then done for each of the steps in a given workflow, e.g. by looking at queue build up and resourcing for QC as shown in Figure 8.

SERVICE GOVERNANCE

Figure 1 illustrates our general approach to service governance. Specifically now we imagine a scenario in which content producers and consumers use an 'archive service' (sitting in the 'application channel'). This may be local or remote to the users and may simply be another department of their organisation or may be outsourced. To use the service, there must be an offline negotiation process in which the terms of a contract including the precise SLA terms are determined and agreed between the service users and the archive, resulting in a contract. Some of these SLA terms are encoded into an electronic SLA. The service management systems are used by archive managers to monitor and enforce the SLA terms and to manage the resources required to provision the archive service.

We have developed a general service management system and an integrated data service. Content producers and consumers use the data service to ingest and access content. The data service executes ingest and access workflows using clustered computational resources to create content thumbnails, browse proxies of video, extract meta-data and check integrity. It uses an object capability access control mechanism [10] and reports usage, described by metrics, to the management system via an asynchronous protocol. The usage information enables the management system to calculate QoS terms and, by comparing to the deployed policies, trigger configuration updates on the data service to attempt to address deviations in expected behaviour.

This management system is distinct from storage management systems widely used in archiving which apply rules to manage copies of data on multiple storage tiers and systems. Rather, the system discussed here is focussed on measuring the metrics necessary to understand and control the QoS experienced by the users of the service.

Metrics are a means of describing the behaviour of a running system in order to calculate QoS. Each metric is a reference to an OWL concept that relates it to a well-defined metric model expressed as an ontology [11].

The following metrics are reported by the data service:

- The amount of time the service has been in existence (seconds).
- The amount of data ingested (bytes).
- The number of MXF frames ingested.
- The amount of data accessed (bytes).
- The rate of data ingest and access (bytes per second)
- The amount of data stored (bytes).
- The product of the data quantity and the time stored (byte.seconds).
- The number of file corruptions observed.
- The number of MXF frames corrupted (using a tool from RAI).
- The amount of data completely lost (bytes).
- The time from an access request being received to the start of the data being accessed (seconds).

The service management software also aggregates information across all services and SLAs and monitors the whether each data service is 'up' or 'down'. The data service permits the service management software to control the access bandwidth (kB/second) and whether access to the entire service is suspended or not.

Service Level Agreement

Users of the archive service agree an SLA for their use of the service. Terms in the SLA include both service guarantees (e.g. [T1] below) and consumer constraints (e.g. [T2]

below):

responseTimeGuarantee := mean(“http://mserve/responsetime”, perDay) < 5 seconds [T1]

accessLimit := total(“http://mserve/access”, perDay) < 100 GB [T2]

In this case, two metrics reported by the data service are used: the time from receiving an access request to the data download beginning (identified by the URI “http://mserve/responsetime”) and the quantity of data accessed (identified by the URI “http://mserve/access”). The first item defines a new term that the service provider intends to keep: “responseTimeGuarantee” is a Boolean and indicates whether the mean value of the response time metric during the current 24 hour period is less than 5 seconds. The second item defines a term for the customer to keep to: “accessLimit” is a Boolean indicating whether the customer has accessed less than 100GB during the current 24 hour period.

The archive manager can use the service management system to define policies to take automated actions to manage resource levels, to enforce SLA terms and to change the state of the archive services. The policy term [T3] below describes a customer management policy:

accessSpeed := if (total("http://mserve/access", perDay) > 100 GB) { 100 kB/s } [T3]
else { 4000 kB/s }

The expression on the right hand side is evaluated every time additional access is recorded. If the total access is greater than 100GB then it evaluates to 100kB/s otherwise 4000kB/s. The result of this evaluation is assigned to the service property “accessSpeed”. In this way, the archive manager provides the users with a ‘soft’ limit of 100GB data access, above which their access speed is greatly decreased.

Part of the power of the service management system is that it does not have any of the metrics, QoS terms or management terms pre-defined. The metrics are URIs with characteristics defined by an ontology and the QoS and management terms result from configurable mathematical and logical terms.

NEXT STEPS

The previous two sections have discussed both the modelling and the operational monitoring and management of services, but not the link between these two aspects of service governance. We are now working on bringing these two together, feeding monitoring data from the services into the models so that the models can reflect reality more precisely and better inform the operators who define and update the automatic service management policies.

Once the storage model can be synchronised with the state ‘now’ and be parameterised with probability distribution functions generated from historical monitoring records then its utility in answering important ‘what if’ questions will be hugely increased. It would then be possible to ask what the maximum ingest rate that could be sustained by the current system was or what additional resources were required to support a new SLA with an associated usage pattern. In addition, optimisations could be performed, for instance to discover the optimum scrubbing interval, trading off cost and data safety. By linking the model to reality, the answers generated will be directly applicable to operational policy decisions.

Similarly, by monitoring and gathering statistics for the media transfer chain and synchronising the model with the actual system those managing the transfer chain will have a powerful tool for predicting future performance, helping managers assess the impact of changes to the workflow and make better informed decisions.

A final advantage of being able to synchronise the models with the real monitored systems is that the models can be validated and improved by storing predictions and comparing with the current state at a later time.

CONCLUSIONS

We have presented a tool for simulating storage integrity and cost over time, a tool for simulating a media migration workflow and a service management system along with an integrated data service. The simulation tools can already provide useful insights into the complex systems that they model, where resources are limited and trade-offs a necessity. The service management system and data service make up the other layers of the framework, providing automated monitoring and management of customer-facing services. To complete the framework, data from the live systems will be processed and fed into the simulation and modelling tools to synchronise their state with 'now', provide pertinent predictions to operators and validate the models.

ACKNOWLEDGEMENTS

PrestoPrime is an EC supported 7th Framework Programme ICT project (FP7-231161) coordinated by INA (Institut national de l'audiovisuel) in France. Partners include BBC, RAI, ORF, B&G and others. For further information see www.prestoprime.org

REFERENCES

[1] OAIS Blue Book, http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] Addis, M. *et al.* (2010). "Threats to Data Integrity from Use of Large-Scale Data Management Environments," PrestoPRIME Deliverable ID3.2.1,http://www.prestoprime.eu/

[3] Barroso, L. A. and Holze, U. (2009). The Datacenter as a Computer: An introduction to the design of warehouse-scale machines. Google Inc. Synthesis Lectures on Computer Architecture no. 6. Published by Morgan and Claypool.

[4] Moore, R. L., D'Aoust, J., McDonald, R. H. and Minor, D. (2007). Disk and Tape Storage Cost Models. In Archiving 2007

[5] Elerath, J. (2007). Hard Disk Drives: The Good, the Bad and the Ugly!, Queue 5, 6, p28-37. http://doi.acm.org/10.1145/1317394.1317403

[6] Jiang, W. *et al.* (2008) Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. FAST '08. http://www.usenix.org/events/fast08/tech/jiang.html

[7] Addis, M. *et al.* (2010). Audiovisual Preservation Strategies, Data Models and Value-Chains. PrestoPRIME Deliverable D2.2.1 http://www.prestoprime.eu/

[8] Addis, M. *et al* (2011). "Digital Preservation Strategies: the cost of risk of loss for AV Content". Jan/Feb 2011 edition of the Motion Imaging Journal of the Society of Motion Picture and Television Engineers (SMPTE).

[9] Cunningham, S. and de Nier, P. (2007) File-Based Production: Making it Work in Practice," http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP155.pdf

[10] Miller, M.S., Yee K. and Shapiro, J.S. (2003) Capability Myths Demolished. Technical Report SRL2003-02. Systems Research Lab, Johns Hopkins University.

[11] Surridge, M., Chakravarthy, A., Bashevoy, M. and Hall-May, M. (2010) Serscis-Ont: A Formal Metrics Model for Adaptive Service Oriented Frameworks. Second International Conference on Adaptive and Self-adaptive Systems and Applications