A Service Oriented Approach to Online Digital Audiovisual Archives

Matthew Addis, Richard Lowe, Charlotte Norlund,

University of Southampton IT Innovation Centre, Southampton, UK

E-mail: {rl, ccfn, mja}@it-innovation.soton.ac.uk

Abstract: **In many parts of the audiovisual community the boundaries between the environments used for content creation, distribution and archiving are becoming blurred. A transformation in the way that electronic media is created and consumed is being followed by a transformation in the way that this content is archived, repurposed and reused. Traditionally, archives sit at the tail end of the content lifecycle and provide a place where content 'ends-up' for safe keeping. However, digital audiovisual archives are now increasingly 'embedded' as active facilities within wider networked infrastructures and content-centric processes. The archive becomes an integrated repository of audiovisual assets which are under continuous development and reuse. This paper presents work done in the UK AVATAR-m project on service-oriented approaches to digital permanence and preservation of audiovisual content. In particular, we recognise that the business models and processes surrounding the storage, preservation and access to digital assets are evolving fast and transcend traditional organisational boundaries. Storage and access to archive content now takes place across organisational boundaries and there is a nascent but growing market for outsourced archive hosting. Our approach embraces this new world where archives can be both deployed in-house and as third-party services. Our specific focus is how to specify and then govern federated storage services in a way that ensures the long term safety, security and accessibility of audiovisual assets in a managed and cost effective way.**

Keywords: Digital archiving, digital storage, service oriented architectures, audiovisual, preservation, archive management

1 OVERVIEW

AVATAR-m is a UK collaborative R&D project supported by the Technology Strategy Board where the IT Innovation Centre, BBC, Xyratex and Ovation Data Services are developing an innovative approach to large-scale long-term digital archiving within distributed storage infrastructures. This paper presents work from the project on tools to support the planning and management of service-oriented data archiving infrastructures. A key feature of our approach is the recognition that archiving is an integral part of content production, distribution and consumption processes. The use of service oriented models, including the delivery of archive hosting through third-party services, provides the key to integrating archiving activities into wider media-centric environments in a way that still allows the archivist to achieve their primary mission – the safety and longevity of their assets. Our tools allow content-centric workflows within an organisation to be analysed in order to profile the generation and consumption of archive assets including the requirements for safety, security, longevity and accessibility. These profiles then allow storage provision to be planned in terms of long-term access, ingest and retention and technical specifications created ready to be matched against storage solutions or managed services.

We also discuss how service oriented architectures using automated policies and service level agreements can be used to deliver online archive functions in a managed way within an enterprise, when outsourcing archive hosting, or when collaborating with external organisations.

2 BACKGROUND

Current projections are that over 90% of all new information is digital and that the volumes generated over the next two years will be larger than the total volume of all information ever created previously in human history.

This is as true for audiovisual content as it is for other types of digital information. For example, YouTube is growing at about 20 percent every month, which equates to over 300% per year. In the professional audiovisual (AV) archive world, UNESCO estimates there are 100M hours of content in existence. Broadcast archives project that this will grow at 5M new hours every year. Considering the storage requirements of today's higher resolutions and frame rates, means that the physical volume of some archives will double within as little as 18 months.

Therefore, when considering all the new devices, techniques, services and business models with which to create, distribute and enjoy audiovisual media, we must also consider how this content will be archived and how it can be maintained in a way that allows it to be easily accessed and used for years to come. The creation, consumption and archiving of audiovisual content are inseparable topics, yet archiving tends to be s an 'after thought' and is often upstaged by the glamour of new forms of content and the experiences they bring.

We already exist in a world where the systems used to store 'contemporary' content are starting to dwarf those used to store 'archive' content. Yet it is also true that archive content is being re-used more often and there is an ever stronger need to integrate archive content into production and distribution processes. The concept of an archive holding the unique original under lock and key in a separate environment has lost its meaning.

Corresponding author: Matthew Addis, IT Innovation Centre, 2 Venture Road, Chilworth, Southampton, UK, + 44 (0) 23 80 760 834

We are now in an era of direct archive integration into production, distribution and consumption workflows, with dynamic preservation processes required as a consequence. For example, the BBC Digital Media Initiative project [1] aims to deploy a completely tape-less environment across the whole organisation over the next 5 years, which includes the archive as an integral part and allows seamless working with externals such as independents and post production houses.

This paper examines in more detail the issues surrounding this new way of working and presents some of the techniques and technologies developed by the AVATAR-m to address these issues.

3 STATE OF THE ART

Whilst there is intensive interest in preservation strategies for digital content [2][3][4][5], in general there is little work on practical implementations tailored for the needs of audiovisual content.

Audiovisual content presents demanding challenges for digital preservation, especially given the preservation ideal of storing content uncompressed. Standard Definition digital video has an uncompressed data rate of about 270 MBit/s and even when stored with compression, e.g. 50MBit/s DV, multiple Petabytes of storage are required for a typical broadcast archive. HD requires five times as much space. In digital cinema, 4K requires up to 30 times the data rate of SD and for 3D cinema with twin data streams at up to 144 fps the volumes are truly vast. This presents a real problem, not least the cost, where estimates range from 'half the price of analogue' [7] to nearly 'twelve times higher' [8].

For example, the OAIS Reference Model [6] defines some of the processes required for long-term preservation and access to information objects, but does not specify how to monitor audiovisual objects or the systems they are stored in, identify when migration should take place or to what an audiovisual object should be migrated to.

More widely, different archive implementation models need to be considered including value chains and business models delivered through multiple service providers or organisations (e.g. outsourced services, federated preservation across organisations etc.). These value-chains and business models are liable to evolve rapidly over time because of the relative rates at which storage, networking, processing are evolving [9], e.g. as evidenced by the explosion in online services such as Amazon S3, EC2 and SQS [10].

The economies of scale, power, cooling and staff costs that can be achieved by organisations like Google [11], mean that as network costs continue to fall, in-house solutions will become increasingly expensive compared with outsourced or federated models. Different approaches will be applicable depending on the type and volume of content or the need for access across organisational boundaries, and the use of mixed models is likely considering robust preservation strategies typically involve multiple copies of content in multiple locations to mitigate against technical obsolescence or content loss.

Whilst audiovisual archives typically use dedicated in-house systems for storage and processing (e.g. transcoding) of their assets, various technologies exist to support data federation and remote data services in distributed environments. Many have emerged from the Grid community, including storage services and high-performance data transfer tools, e.g. GridFTP[12], SRB[13] and RFT[14]. These are used as part of Data Grid Management Systems[15] to support the needs of large-scale scientific applications e.g. High Energy Physics Experiments at the CERN LHC. iRODS[16] is one that has already been used for digital library applications, persistent archiving, and real-time data systems, where management policies (sets of assertions that these communities make about their collections) are characterised in terms of rules and state information. Remote access to archive hosting services is yet to emerge in the broadcast industry, although there are services for remote access to data for distribution, e.g. VIIA from Ascent Media[17] and data transfer within the enterprise, e.g. DIVAGrid from Front Porch Digital [18].

Critical to provision of services in trusted archive environments is the use of policy-based service governance, which is based on two principles: that the non-functional aspects of a service including performance should be agreed in a service level agreement (SLA), and that the service should be managed, preferably in an automated (self-governing) management environment, so that it conforms to its SLA. Initiatives to standardise the way SLAs are made and represented includes WSLA[19] and WS-Agreement[20] from the Open Grid Forum (OGF). This provides a high-level structure for an agreement on the quality of service (QoS) offered by a service provider to a consumer, plus simple protocols for establishing and monitoring such agreements. Web Service based infrastructures with explicit support for automated service management using policy-driven SLAs and QoS include FP6 NextGRID [21], FP6 TrustCOM[22] and IT Innovation's GRIA[23] technology. These projects recognise that trust and security (e.g. to support assertions of integrity and authenticity) is equally important in distributed environments, e.g. NextGRID work on interoperation across heterogeneous security environments, including X.509, SAML and Kerberos token exchange.

The digital library community has meanwhile been busy creating software frameworks for implementing preservation environments. These include open source solutions, e.g. DSpace[24] which provides standard services for ingestion and access and is ported to run on top of SRB for managing distributed data, Fedora[25] which associates display functions with each data type, allows relationships to be imposed on records, and maps semantic labels on records to an ontology, as well as simple, off the shelf systems such as Greenstone[26], and commercial systems, e.g. ExLibris[27] – however none are designed specifically with the challenges of AV content in mind.

These specific developments are converging through wider work in the Service Oriented Architecture (SOA) community where Web Service standards, e.g WS-Interoperability[28], provide a base technology for distributed services. This approach is being used by a new generation of inter-organisation production and post-production infrastructures, e.g. in MUPPITS[29], PRISM[30] and BeInGrid[31] as well as products such as Signiant's Digital Media Distribution Management Suite[32]. Current focus of the SOA community is open specifications for the management layer building on the work from WSDM/WSRF and WS-Management, along with service orchestration, e.g. using XPDL[33], ebXML[34], and WSBPEL[35]. Whilst workflow technology in general is widely used in production and post production, e.g. Autodesk workflow products [36], the use of workflow standards and techniques is a current topic of discussion in digital preservation [37], and specific tools are only just emerging, e.g. PAWN (Producer Archive Workflow Network) [38].

In summary, there are clear indicators that digital archiving is changing rapidly in the AV community and new business models can be anticipated based on archive service provision. However, the technology state of the art is one of fragmentation where individual communities, e.g. the broadcast industry, digital libraries, and SOA, each provide pieces of the puzzle. The challenge is one of integration and adaptation to the specific challenges of audiovisual content.

4 APPROACH

Our solution is based on three core components. Firstly, we use GRIA and aggregated storage as the basis of secure and managed archive hosting services that operate across administrative domains and can be federated with internal systems. The OAIS standard is used to specify the interface of these archive services. Secondly, we use a multi-level model of archive requirements to allow the concerns of the archive manager (assets, users, safety, longevity, value) to be separated from the specifics of a particular technical solution (disk, tape, networking etc.). This is done through profiles for archive ingest, access and retention which specify what goes in and out of an archive. Thirdly, we are developing simulation and modelling techniques to analyse content-centric workflows to determine the workloads these place on an archive and the variations that are likely to occur on a range of timescales.

Storage in AVATAR-m is heterogeneous, reflecting the broad range of storage types that an archive may typically utilise. The emphasis in our solution is on networked storage, such as spinning disk or media jukeboxes, which may or may not be configured within a SAN or NAS. Additionally, online remote storage provided as a service is also supported to allow archives to make use of third-party storage services such as Amazon. Our approach is to combine these disparate storage types and locations, so they are aggregated together into a single storage solution as shown in Figure 1.

Adapters are used for each storage type that the storage aggregator interfaces to, but since most operations are done at the file system level additional adapters are only required for storage services, which offer different APIs. Rather than assigning each asset to a specific tier, available storage locations are ranked dynamically using a cost function and multi-objective optimisation based on factors such as the current and average read/write rates and availability. The use of the storage is also monitored ensuring that content that is accessed frequently is made available from higher-ranked (and therefore faster) locations, whilst content that is not accessed often is moved to slower storage. The rules that determine what gets moved can be modified through management policies that can be assigned to specific items or classes of items, such as all files of a certain type or belonging to a certain user or project. This is similar to hierarchical storage management (HSM) systems, but with the advantage in our case of being able to utilise third-party storage services as well. Interactions with the aggregated storage happen through GRIA (Figure 2). GRIA is an open-source service-oriented infrastructure (SOI) designed to support B2B collaborations through service provision across organisational boundaries in a secure, interoperable and flexible manner.

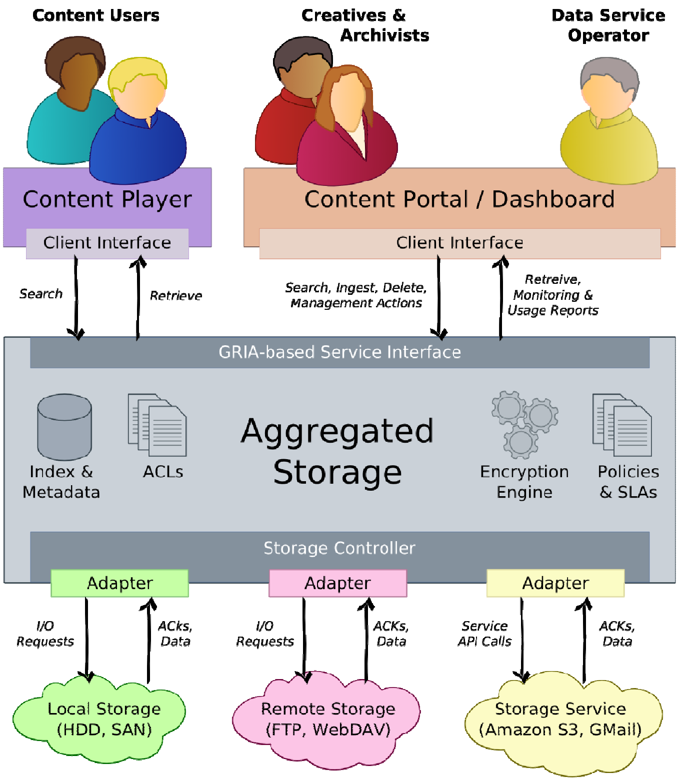


Figure 1: AVATAR-m aggregated storage

GRIA makes use of business models, processes and semantics to allow service providers and users to discover each other and negotiate terms for access to archive services. Service providers and customers trade resources (applications, data, processing, storage) under the terms of bilateral SLAs which describes quality of service (QoS) and gives a promise to provide services, for instance to store and provide access to data for a particular period of time.

In our model, ingest will use the OAIS model, i.e. a Provider uploads a Submission Information Package (SIP) to the Service Provider through a Data Submission Session. The SIP includes the content and preservation information (e.g. the retention schedule). Likewise, content access will also use the OAIS model, i.e. a Consumer downloads a Dissemination Information Package from an OAIS service provider through a Data Dissemination Session.

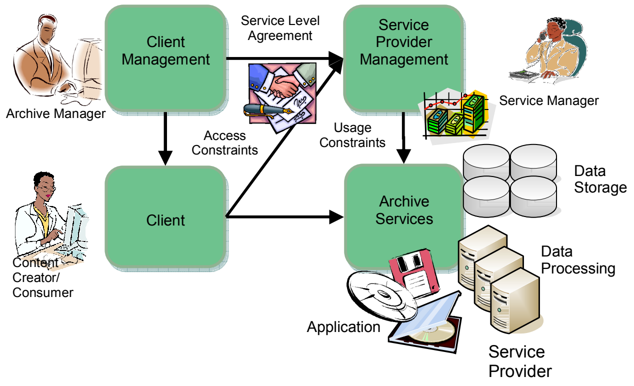


Figure 2: GRIA service oriented framework in AVATAR-m

Management of Ingest and Access is then done by GRIA according to the SLAs. This is essential for preservation activities using storage services since they need to run efficiently and dependably so the content is not subject to unnecessary risk. This is done through the storage adapters, where instrumented storage and data transfer systems communicate data-centric metrics, e.g. I/O (max, min, average), storage usage, frequency of access, latency etc. Control points allow GRIA to react to this information to manage the services, e.g. stopping access or upload, throttling bandwidth, or giving different users priority over each. Rules and policies within GRIA encapsulate how to go from the reported metrics through a series of decision points that invoke these control points, e.g. to limit the volume of content submitted each month according to the agreed terms of the service. WS-Security is used as the basis of security and GRIA provides both transport (SSL) and message-level security (X509 or SAML). The use of SAML tokens allows GRIA to federate security policies between domains using WS-Federation patterns. This can be integrated with local security management, e.g. LDAP or Active Directory, at the client and server sides, to allow dynamic and automatic access control between organisations. For example, a content owner could set a policy of who can access their content, including people in other organisations that they trust, and this can be dynamically and automatically propagated to the access control mechanisms used for data delivery.

5 RESULTS

Whilst our solution is still under development, experience with an initial implementation has found that there are often significant differences between the parameters with which storage services are defined (storage capacity, access latency, delivery bandwidth etc.) and the level at which archive operators characterise their archive (rates and volumes for ingest and access, retention scheduling to encapsulate value, preservation priorities and asset safety). To address these differences, we developed a storage planning tool (Figure 3) that allows archive requirements to be specified using parameters (e.g. data volumes and data i/o) that are both application and technology implementation neutral. The tool can be used by an archivist, external service provider, or in-house IT manager to define SLAs in archivist terms or to interpret resource-level SLAs.

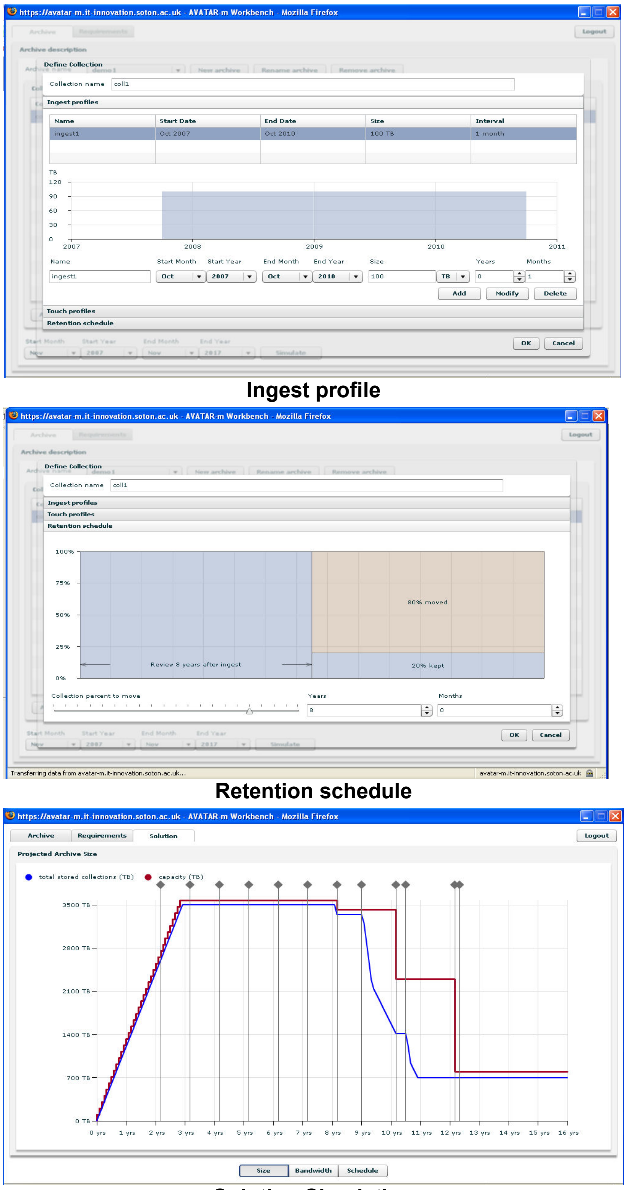
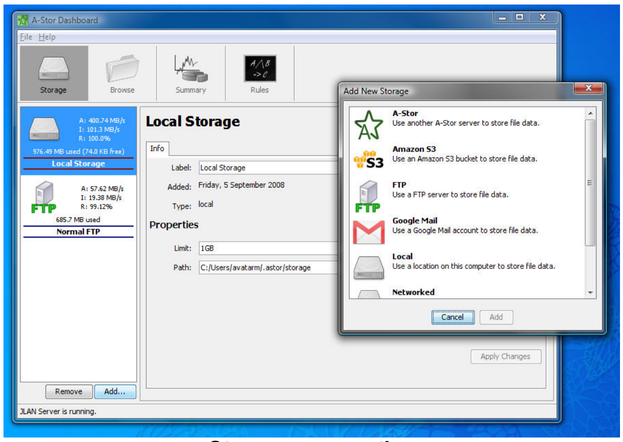


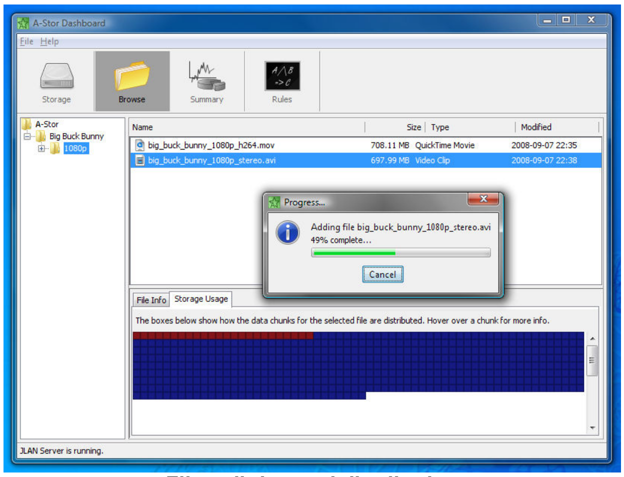
Figure 3: Archive requirement specification using the Storage Planning Tool

Through a series of screens, the user can specify one or more collections of assets and the associated ingest, access and retention profiles. For example, a collection might be born digital content of a particular genre or it might be a particular type of analogue carrier being migrated into digital form in a preservation project. The ingest profile specifies the rate at which items are put into the archive and can be expressed in various ways, e.g. items per month or terabytes per year. The access profile specifies how often material is likely to be accessed and can be expressed as an average rate or as a periodic activity. The retention schedule specifies how long each item of content needs to be retained before it is re-appraised and includes an estimate of how much content is likely to be retained after that point. Ingest, access and retention profiles are aggregated across the collections to define the overall needs of the archive. The tool allows simple storage solutions to be simulated (e.g. tape libraries) using technology roadmaps (e.g. LTO data tape) to profile investment and migration and find deviations from the archive needs, e.g. resulting from device contention during concurrent migration and access.

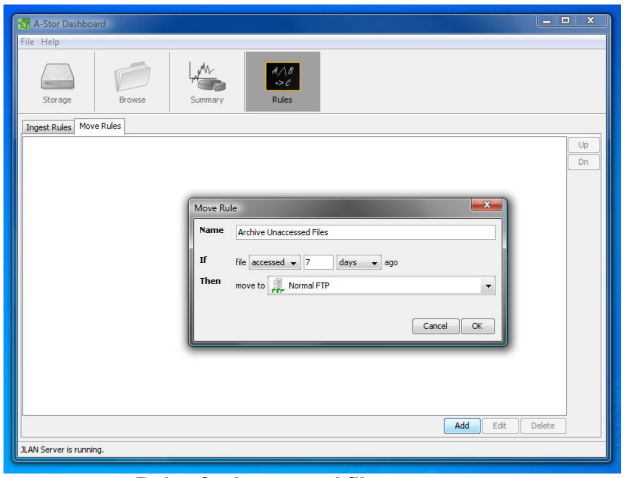
When implementing the framework (see Figure 4) to allow files to be stored on one or more storage locations, e.g. local disks or remote storage as a service, we found that several features were essential. Firstly, archive content needs to be as easy to use as local files. Therefore, irrespective of physical file location, files are accessed as if they are local, e.g. they can be exposed as a mapped drive in Windows. Transfer from wherever the file is actually stored in the archive is completely transparent. Secondly, it needs to be easy from a management point of view to use both local and remote storage in the same way, including remote storage provided as a service by a third-party, e.g. Amazon S3. Therefore, a generic set of parameters (capacity, bandwidth, availability) is used to describe all storage types and allow them to be ranked and managed as a single unified list of storage locations. Thirdly, we found that a flexible set of rules is needed to define how content is distributed across or moved between one or more storage locations. These allow automated policies to be applied that encode the business rules of the archive for safety, accessibility, cost, security etc. An essential feature is the ability to chop up files into 'chunks' and use rules to define what happens to those chunks, e.g. replicate the most important chunks multiple times for safety, keep the index or metadata chunks of a file on fast access storage for quick interrogation, ensure confidential file chunks only get stored in certain places where security is sufficient and so on. Currently, we chop files up into equal size chunks and have simple ingest and move rules, e.g. to distribute the chunks across storage according to storage performance or to move files between tiers of storage based on file size or access frequency. However, the architecture we use allows plug-ins to be added that could understand files at an application level and then decide how best to chunk these files, e.g. by disassembling an MXF asset into component pieces, extracting key frames from an MPEG stream and so on.



Storage aggregation



File splitting and distribution



Rules for ingest and file movement

Figure 4: Storage management dashboard

6 FUTURE WORK

In the next phase of the project, we plan to develop a combination of process modelling and statistical techniques to calculate the workloads placed on an archive from the processes that involve the archive, including ingest, access, transcoding and maintenance (e.g. through migration). This will combine workflow specification languages and enactment engines, queuing theory, and Monte Carlo simulation techniques to analyse the variability of archive workloads and hence the flexibility needed in the systems used to implement the archive.

The use of more advanced requirements estimation will form the basis of round-trip capacity planning, SLA definition, archive service provisioning, and service usage auditing and reporting, including the case where archive hosting is outsourced. We expect to extend our use of automated archive rules, for example so we can support policies for who can access what. We also plan to allow rules to be applied to different types of content, e.g. for transcoding video content on ingest to create proxies for access, or to be triggered in reaction to system events so we can react to drops in availability of storage locations, e.g. if a copy is on Amazon S3 and it goes offline then replicate one of the remaining copies to another location to maintain safety.

Our ultimate objective is to demonstrate a decision support tool (dashboard) for planning, monitoring and managing archiving using distributed storage infrastructures in a way that allows suitability, flexibility, scalability and cost to be investigated, trade-offs to be explored, and best-fit solutions to be chosen from the perspective of both the consumer of the services and the provider of the services.

7 CONCLUSIONS

In this paper we argue that audiovisual archiving is becoming an integral part of content production, distribution and consumption processes. The use of service oriented models, including the delivery of archive hosting through third-party services, provides the key to integrating archiving activities into wider media-centric environments in a way that still allows the archivist to achieve their primary mission – the safety and longevity of their assets.

A move towards service-oriented and federated archive systems brings with it several challenges. There is a need for archive managers to communicate the requirements of the archive to the technical implementers of archive systems whether in-house or outsourced. There is a need for tools supporting capacity planning over long timescales to ensure the IT systems are sufficiently scalable but also planning with fine granularity to ensure systems are robust and flexible to peak loads. Finally, archives as services need to be embedded within content-centric environments and deployed across administrative domains with well defined and automatically managed SLAs and QoS specifications.

AVATAR-m addresses these challenges through the use of aggregated and federated storage, a service oriented infrastructure to access and manage this storage, and user interface tools to help with capacity planning and decision support. This allows archive owners to concentrate on the long term management of their content in a secure, safe, and cost effective manner.

Acknowledgements

AVATAR-m is a £3M UK Technology Strategy Board supported collaborative R&D project that started November 2006 and runs until 2010. The project involves the BBC, Ovation Data Services, Xyratex and the IT Innovation Centre. For more information please contact avatar-m@it-innovation.soton.ac.uk

References

[1] http://backstage.bbc.co.uk/news/archives/2007/09/bbc_dmi_project.html

[2] Digital Preservation Lifecycle Management for Multi-media Collections. Arcot Rajasekar, Reagan Moore, Fran Berman and Brian Schottlaender. Book Chapter. LNCS http://www.springerlink.com/content/p3m497wt5452012u/.

[3] CLIR pub106 Building a National Strategy for Preservation: Issues in Digital Media Archiving http://www.clir.org/pubs/abstract/pub106abst.html

[4] Evaluating preservation strategies for audio and video files. Carl Rauch et al. http://www.ifs.tuwien.ac.at/~andi/publications/pdf/rau_digrep05.pdf

[5] A Foundation for Automatic Digital Preservation, Miguel Ferreira, Ana Alice Baptista and José Carlos Ramalho Ariadne Issue 48 July 2006. http://www.ariadne.ac.uk/issue48/ferreira-et-al/

[6] CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2003. http://public.ccsds.org/publications/archive/650x0b1.pdf

[7] "Archiving Movies in a Digital World". David Cavena et al. SUN Microsystems report, VERSION 2.1, June 8, 2007; 29pp David.Cavena@Sun.COM http://wikis.sun.com/display/SunMediaSpace/2007/11/05/Archiving+Movies+in+a+Digital+World

[8] "The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials" Academy of Motion Picture Arts & Sciences, 2007; 74pp available from the academy: http://www.oscars.org/contact/council.html

[9] For example, see Gilder's law: http://www.netlingo.com/lookup.cfm?term=Gilder's%20Law. Edholm's law: http://ieeexplore.ieee.org/iel5/6/29070/01309810.pdf?arnumber=1309810. Kryder's law: http://www.mattscomputertrends.com/Kryder's.html. Moore's law: http://en.wikipedia.org/wiki/Moore%27s_law

[10] Amazon Web Services: http://www.amazon.com/AWS-home-page-Money/b/ref=sv_hp_4?ie=UTF8&node=3435361

[11] IEEE Spectrum: Google Watches Its Watts http://spectrum.ieee.org/jul07/5382

[12] GridFTP http://www.globus.org/grid_software/data/gridftp.php

[13] Storage Resource Broker http://www.sdsc.edu/srb/index.php/Main_Page

[14] Reliable File Transfer http://www.globus.org/toolkit/docs/4.0/data/rft/

[15] DGMS http://www.sdsc.edu/srb/index.php/DGMS

[16] https://www.irods.org/index.php/IRODS:Data_Grids%2C_Digital_Libraries%2C_Persistent_Archives%2C_and_Real-time_Data_Systems

[17] http://www.ascentmedia.co.uk/

[18] http://www.fpdigital.com/

[19] http://www.research.ibm.com/wsla/

[20] http://www.ogf.org/documents/GFD.107.pdf

[21] FP6 NextGRID Project. http://www.nextgrid.org/

[22] FP6 TrustCOM Project. http://www.eu-trustcom.com/

[23] GRIA, http://www.gria.org

[24] DSpace http://www.dspace.org/

[25] Fedora http://www.fedora.info/

[26] Greenstone http://www.greenstone.org/

[27] ExLibris digital library solution: http://www.exlibrisgroup.com/

[28] WS-Interoperability. http://www.ws-i.org/

[29] Multiple User Post-Production IT Services (MUPPITS) http://www.muppits.org.uk/

[30] PeRvasive Infrastructure and Services for Media (PRISM) http://www.bbc.co.uk/rd/projects/prism/index.shtml

[31] BeInGrid Movie Post Production Workflow Experiment http://www.beingrid.eu/be2.html

[32] http://www.signiant.com/dmdms

[33] Workflow Management Coalition Workflow Standard, XML Process Definition Language (XPDL), Document Number WFMC-TC-1205.

[34] OASIS standard v2.0.4, ebXML Business Process Specification Schema Technical Specification v2.0.4, December 2006.

[35] Alves, A., et al. (Eds), Web Services Business Process Execution Language Version 2.0, OASIS Committee Specification, Jan. 2007.

[36] http://usa.autodesk.com/

[37] See ERPANET WORKSHOP ON WORKFLOW, Budapest, October 13-15 (http://www.erpanet.org/events/2004/budapest/index.php)

[38] PAWN: Producer - Archive Workflow Network in Support of Digital Preservation, Mike Smorul ; Joseph JaJa ; Yang Wang ; Fritz McCall, UMIACS-TR-2004, 2004.