Best Practices for Digital Preservation in Smaller Institutions: *The SCOLA Model*

Introduction:

SCOLA's broadcast preservation methods are a model for institutions archiving video and audio digital information. SCOLA is a § 501 c (3), educational, non-profit staffed by 60 individuals. Even though SCOLA is a small organization, it archives materials on a large scale. SCOLA's Digital Archive is a compilation of searchable and downloadable resources pertinent to foreign language study and is invaluable in its historical significance. Users can find past and present foreign TV broadcasts, machine-generated translations and other important data available in 60-plus languages and dialects. Since 2007, we have preserved 80 terabytes of video media in the Digital Archive.

About SCOLA: specializing in less commonly taught and many other languages, SCOLA is a premier source for language training material. Across the 12 services on the SCOLA site, 150+ countries and 175+ languages and dialects are represented.

I will describe the design of our digital archiving process from the standpoint of the needs and resources of a smaller institution, but one also that strives to remain at the forefront of modern technological applications.

For the purpose of this discussion, I will divide the archiving process into four main components**: *Ingest, Publish, Storage and Use[rs].*** The eight channel (24/7) ***Ingest*** process begins on the SCOLA campus where the format conversion, editing and ***Publishing*** is accomplished. A 100 meg data pipe runs between SCOLA and our IT hosting partner (First National Technology Services) (in Omaha, Nebraska,) where the files are stored in multiple file formats. Two copies of our video broadcasts are created and ***Stored*** – one on Blu-Ray discs on Netherlands-based DAX jukeboxes, and the other copy is physically housed off-site. SCOLA is a web-based subscription service whose end-***Users*** are primarily foreign language learners from the educational sector in addition to linguists from various government agencies.

Ingest

- Video files broadcast and recorded according to schedule have basic metadata generated in the file name: SCOLA Service Information, SCOLA broadcast date, channel, country and language codes.

Publish

- MPEG-2 files are recorded to a server on the SCOLA campus and are automatically saved to a designated, unedited folder on the server. Files are accessed by remote desktop for manual trimming. These trimmed files are dropped into a Rhozet Carbon Coder and converted to Flash [flv], Windows Media [WMV] and MP4 H.264 formats.

- The files next appear on an administrative page, where additional metadata is manually added: original broadcast date, program names and categories. Keywords are authored and tagged manually. Corpus metadata for machine-translated files and transcriptions is automatically indexed and tagged.

Use/End User

- Published files can be searched by using the metadata fields appended in the Ingest process and/or keywords.
- Files can be viewed online as Flash, MP4 H.264 or Windows Media.
- Files can be edited online.
- Files can be downloaded in MPEG-2, Windows Media, or MP4 H.264 formats.

Storage

- Flash, Windows Media, and MP4 H.264 files are permanently stored on the Network Attached Storage [the NAS], while MPEG-2 files are stored only temporarily. After 7-10 days MPEG-2 files are copied to Blu-Ray discs on a DAX Jukebox (which holds approximately 35 terabytes each). A second disc is recorded by the DAX and is physically stored in a secure off-site facility.

II. *A Formula for* **Ad Hoc Ingesting**

Hard media video files are manually processed in ingest stations with multiple DVD and VHS decks, etc. using an application DVD Lab Edit Studio 6. Similar to the automated Ingest process, these files have the same basic metadata tagged and identifiable in the SCOLA file name.

III. Ingesting Learning Objects

Learning Objects [LOs] are multi-genre language lessons – and a SCOLA service – that are based on international television clips from beginning primers to ILR[i] Level 3 content and professionally produced films, which give detailed overviews of various countries. There are two types of Learning Objects that are ingested using the Learning Object Portal on the SCOLA web site: those that are created by a third party and those that are created using the Authoring Tool provided in the Learning Objects Portal.

When a contracted third party creates a learning object, they use their local system. The object is either a flash file or a php file – the latter being a server side scripting language. When the file is ready to be transferred, it is uploaded via the SCOLA Learning Objects Portal. At this point, the creator of the file is allowed to enter the *Name, Title, Description, Language, Topic, Duration, Language Level and any additional tags.* The Learning Objects Portal transfers the file from the administrator's system to the LO server at FNTS, and records the location of the

Learning Object and the attached metadata in the Learning Objects database. The media files embedded in the larger file are extracted and stored on SCOLA's media server.

The LO files are checked for quality control and edited, if necessary, at SCOLA. At this point, the LO is published (changing its status in the database), which allows subscribers to view the LO when they access the Learning Object Portal via the SCOLA website.

When the **Authoring Tool** is used to create a Learning Object, the file content is stored in the Learning Objects Portal Database and the attached media files for the lesson – resources, glossary, questions – are stored on the SCOLA media server. The author attaches the pertinent metadata etc., which is stored respectively in the LO database or on the SCOLA media server. The protocol for quality control, editing and publishing follow the same Ingest process as the creation of third party Learning Objects.

IV. SCOLA's Digital Archiving Project

Envisioned as an educational tool for research and language studies, SCOLA's archive with its ancillary functionality (search engines, translation applications, etc.) is invaluable for those teachers, students, linguists, and general users wishing to find specific programming from a breadth of countries in a host of languages. The archiving project has a multidimensional purpose, and provides affiliates with SCOLA's historical programming, whose availability will allow individuals or groups a greater understanding of events, people and culture.

Within the Archive **Machine translation – Speech-to-Text –** is the automatic translation of textual data by either statistical analysis or dictionary queries. SCOLA's application is a statistical model. Upon ingestation the system processes selected archive media files for Speech-to-Text extraction based on language. Current available Speech-to-Text languages in the Digital Archive are: Arabic, Farsi, French, Hindi, Indonesian, Mandarin Chinese, Pashto, Russian, Spanish and Urdu. Machine based translations are also keyword searchable. The company BBN Technologies is SCOLA's vendor for this protocol.

Speech Phonetic indexing processes the audio portion of each media file, creating a **phoneme** index for subsequent search operations. This allows SCOLA users to search for media file content based on the audio content, not just the associated metadata. Phonetic indexing proves quite useful in searching multilingual media files and various media files with specific content not specified as metadata. Arabic, Dari, Farsi, Korean, Mandarin Chinese and Pashto are presently the languages with this capability. SCOLA's vendor is Nexidia, and we are using their product, NexMiner.

V. The Vision

The idea of a Digital Archive was conceived by SCOLA's president, Francis Lajba. SCOLA was broadcasting thousands of hours of foreign newscasts – "history in the making" – however, the material was not being archived. Francis' philosophy opposed the idea of "just letting these historical artifacts disappear," and believed that if we had the technology to save these materials SCOLA should do so.

SCOLA hired an outside consulting firm to assist in developing the system requirements for a Digital Video/Audio archiving system. The system was required to utilize primarily Commercial-Off-The-Shelf (COTS) based solutions for collecting, processing, storing and distributing audio and video broadcasts.

The objective of the Digital Video/Audio Archiving project was to develop a permanent digital archive of selected programming as needed by the Library of Congress[ii] and the Department of Defense. From a historical perspective the archive was created to preserve valuable television programming as representing globally the cultures of many nations.

VI. Key Design Drivers for the SCOLA Model

Scalability

One of the most important design drivers for the archive project is that the asset archive be scalable in order to support growth of the collection and the many anticipated users who access the digitized collection.

Reliability

The system software and hardware components were chosen in an effort to achieve the most reliable and scalable system. Growth of the SCOLA customer base continues and necessitates flexibility for repurposing collected digital assets. The system component recommendations took into account future growth and its impact on scalability and reliability in a cost-effective manner.

Automation

The architecture reflects additions to the SCOLA infrastructure which are intended to introduce digitization of the collection and process automation. Existing capabilities were taken into consideration and we endeavored to specify components that were either already in use or ones that would be compatible with new capabilities wherever possible. Functionally, these hardware and software elements added additional automation to the workflow processes, and support growth and diversification of our customer base.

Ease of Use

The system software and hardware components were selected, in part, based on usability and, in part, open system architecture, allowing features to be added in the future as needed. Automation and software user interfaces are employed to minimize manual hardware switching making it easier for operators to schedule, ingest, and rebroadcast video content by satellite as well as streaming on the internet.

Cost

We have taken into consideration both existing and desired SCOLA capabilities in proposing the system architecture, weighing costs and benefits when evaluating vendor solutions in order to create an efficient and cost-effective system. While SCOLA is not a large broadcasting

enterprise, various qualitative and quantitative factors necessitate a commercial grade solution of the caliber used by larger broadcasters:

- large quantity and anticipated growth of programs collected and rebroadcast
- scale of the organizational network of contributors and partners
- technical mix and nature of the satellite-based and Internet-based operations
- additional availability of programming to customers from search capability (video-on-demand)
- importance of the mission of SCOLA and its customers

Implementation of such a system enables the SCOLA organization to effectively process the scope of the programming contributions and enables SCOLA customers to readily access the collection. As budgets allow, enhanced capabilities can be easily integrated to keep step with industry trends and technology advancements.

Integration Emphasis

All system software and hardware components have been selected based on the ease of integration, relying on integration of engineered broadcasting hardware and software with computing based information technology. The ease of integration was a highly considered factor while evaluating each component to keep cost estimates accurate and within budget.

Perpetual Storage

Initially, SCOLA was required to collect and digitize programming assets for a period of 6 months due to contractual obligations. It was envisioned that this data retention policy could be extended beyond this time period in the future, raising the need for nearline storage. By utilizing both online storage and cost-effective nearline storage[iii], video content from the permanent digital archive – theoretically – would be saved indefinitely.

Flexibility

The design of the SCOLA digital archive supports various uses for the collected programming including broadcast play out, web-based Video-on-Demand, and the ability to repurpose the assets. Whenever possible, we have avoided proprietary storage schemes and formats in favor of network addressable storage solutions that readily support transcoding and ensure that the media files are accessible — *as much as possible* – by non-proprietary software.

Due to the emphasis on flexibility of the system design, assets can be repurposed for a multitude of projects, post ingest processing functionality including machine translation, and search and retrieval of audio/video assets.

Extensibility

The system lends itself to extensibility because of the design choices to keep the system standards-based, modular and loosely-coupled. Consequently, as the system grows and

additional features are required, the SCOLA Digital Archive has been able to be extended to support these new requirements. This is an important design consideration since most systems evolve and expand over time.

Compatibility

The compatibility and integration of the various vendor solutions is also a critical part of the design since few vendors provide complete end-to-end capability (mostly limited to broadcast only capability), and few provide adequate capability in all aspects of SCOLA's workflow for digital media asset processing. For example, the functionality desired by SCOLA is not simply broadcasting-based, so vendor specifications must represent a "best-of-breed" approach.

Open Standards

Where applicable, SCOLA makes use of open standards recognized by international industry professional and government organizations serving the broadcast industry, networking and computing industry, and those electrical and information standards. For example, use of Motion Pictures Experts Group MPEG standards are supported by the ISO/IEC standards bodies. The W3C administers the standards for XML, SOAP, and WSDL used for web services and component integration. The UTF8 standard supported by the ISO/IEC standards bodies is used for multilingual support. Several system components are Java Platform Enterprise Edition (J2EE)-based which would lends itself to being extensible. Common Internet File System (CIFS) is used for file access across disparate systems in a network without regard to the underlying operating system. The use of open standards ensures support, reusability, portability, and extensibility, providing the basis for future system enhancements and repurposing of digital assets from the permanent digital archive.

Archive Quality

Another key consideration for digital asset storage is ensuring that the digitized format of each captured program provides for the potential transcoding to other digital formats and resolutions. Capture and digitization at broadcast quality format and resolution provide the flexibility to dynamically produce or transcode a lower resolution proxy for quick viewing, a streaming version of the program for Video-on-Demand [VOD], or other format conversions. The hallmark of SCOLA's archive format is the efficiency of storage combined with the ability to use it as a basis for transcoding to other formats.

The recommendation for storage of digital media in the SCOLA system must necessarily take into consideration a reduction of media costs due to the enormous scale of the collection. Retaining the best quality assets necessitates larger sized files. Since the quality and flexibility of each digitized asset is affected by the digital encoding format chosen for archiving, the overall capacity and growth potential of the collection is necessarily affected by cost. At the same time, dynamically transcoding upon demand (either from VOD or scheduled rebroadcasts of the programming) is potentially costly since providing enough capacity to handle the

anticipated customer demand from VOD comes in the form of additional transcoding servers and software instances, which necessitates additional hardware and software licensing costs.

Various potential archive quality digital formats were evaluated, and SCOLA chose (in 2006) the MPEG-2 for several reasons. Primarily, it gives SCOLA the flexibility to transcode the collected programming to other formats. In addition, we try to consider the format preferences of the various organizations served by SCOLA. We also felt MPEG-2 was a good choice since the Library of Congress was a primary SCOLA customer and its representatives had shown support or preference for open standards from the MPEG organization. The MPEG-2 format provides the broad foundation for transcoding to other formats, such MP4 H.264, Windows Media and Flash. However, presently there is an ongoing discussion at SCOLA regarding whether or not we will continue to archive in the MPEG-2 format in favor of the highly improved MP4 H.264 format.

There are also many professional digital video (DV) formats in use in the industry. Whereas digital formats commonly used in the production broadcasting industry could have been used for the archive storage, they generally require a great deal more storage space, which drives up costs. But costs are centered not only in the storage cost; the processing costs when working with high-grade broadcast DV are also cost components.

Broadcast Focus

SCOLA's focus on the broadcast medium influenced a number of the architectural recommendations. This is seen beyond the broadcast subsystem in the collection, ingest, and application subsystems. The broadcast and educational nature of the SCOLA system when combined with the architectural recommendations contained the functional aspects of the system sufficient to support future capabilities, as were previously mentioned, foreign language Speech-to-Text processing and phonetic searching.

VII. Future Considerations

- Obsolescence Prevention
- Scalability (content/functionality)
- Advanced Search Utilities

VIII. Conclusion

At present, SCOLA's web site will be undergoing a major renovation of scale and functionality. Where previously each service had to be searched individually for content result, in the future search capability will be possible across the whole of the site simultaneously, while maintaining the identity of all the SCOLA services.

Adaptability is one of the key components to SCOLA services. Our mission statement reads, in part, that SCOLA emphasizes the importance and effectiveness of modern information technology as a tool in overcoming barriers to global understanding and will remain at the forefront of its application. To this point, reviewing Cloud technology as relates to SCOLA's archival process is a subject we have begun to explore.

Additionally, we are looking at upgrading our video standards. Currently our standards are approximately 400 kbps for videos (mpeg-4, wmv, flv), and certain files are archived from a SCOLA service, World TV Online (WTO), a six day, 24/7, rolling archive, as mpeg-2 at 5 mbps. We are researching and considering changing our specifications in order to have two bitrates.

- 400 kbps – low resolution (mpeg-4, wmv, flv)
- 2 mbps – high resolution (mpeg-4, wmv, flv)

SCOLA is in the process of adapting mobile technology to each of our services, which has been a challenge due to a number of factors, for instance, file size. Nonetheless, the challenge in creating mobile apps has been an extremely interesting one.

Having huge repositories of foreign broadcasts accessible to affiliates has helped SCOLA pursue funding for further research and development of programming and new practices. I hope the discussion that follows opens the door to additional ideas for use by smaller institutions as we strive to survive in this nebulous economic times.

---

[i] ILR: The Interagency Language Roundtable scale is a set of descriptions of abilities to communicate in a language.

Professional working proficiency is the third level in the scale and is what is usually used to measure how many people in the world know a given language

[ii] Our original partnership with NDIIPP stemmed from SCOLA's role as a nexus of foreign news broadcasts, and was part of the Library of Congress' "initiative to collect and preserve important digital information."

[iii] As defined in ***Wikipedia*, Nearline storage** (where the word "nearline" is a contraction of **near**-online) is a term used in computer science to describe an intermediate type of data storage that represents a compromise between online storage (supporting frequent, very rapid access to data) and offline storage/archiving (used for backups or long-term storage, with infrequent access to data).