StorExcel White Paper:

Metadata Fusion

By Kurt Clawson (VP, Technology M&E)

TABLE OF CONTENTS

ABSTRACT

In todays digitally rich computing environments, understanding more about the contents of large data sets consisting of video, image and audio files can be overwhelming but required. Automating the collection and understanding of the associated data sets through metadata tagging can save time and money, and in some cases enhance and accelerate the monetization of digital assets. The key to overcoming these challenges can be as easy as combining commercially available off the shelf technologies, along with an expert knowledge of how to integrate them together in a unified platform that can be leveraged to improve metadata search relevancy and data set understanding. This platform automates the processing and fusion of time-based metadata from traditional media logging, video content analysis, and audio content searching to enhance metadata and provide contextually ranked search results based on feedback from three disparate sources against the original search criteria. Utilizing an automated solution that identifies the complimentary time-based metadata points with the collected metadata then fusing them together through a centralized asset management solution; a more relevant and sophisticated ranking of the search results are displayed in a unified contextual view for more efficient discovery and utilization of assets.

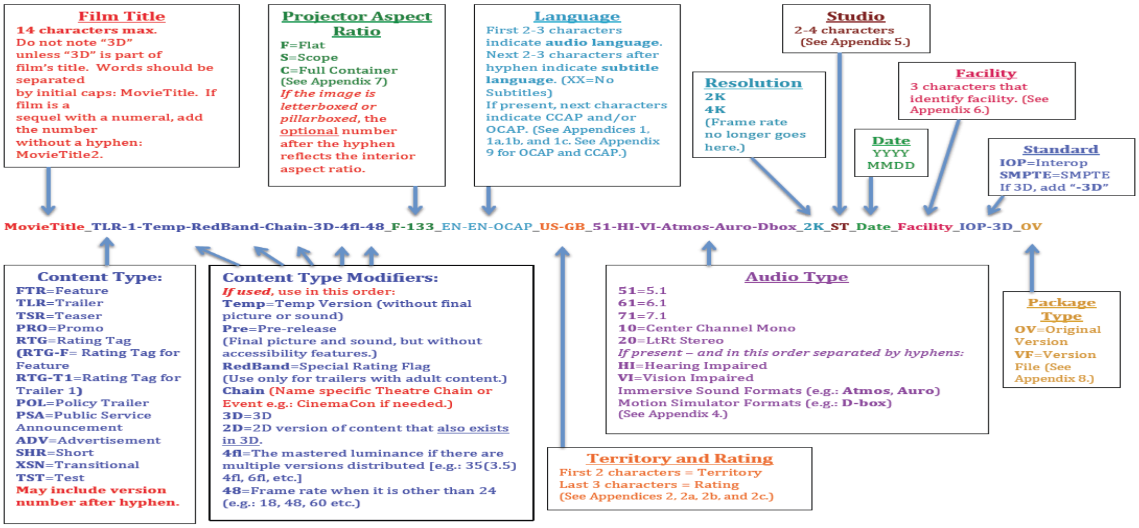
1. THE INTRODUCTION

Before we define our interpretation of "Metadata Fusion", let's first start with the stand-alone definitions of metadata and fusion:

met·a·da·ta (*noun*) – a set of data that describes and gives information about other data.

fu·sion (*noun*) – a merging of diverse, distinct, or separate elements into a unified whole.

In a file-based environment minimal conventional metadata is often available through a number of relatively simple mechanisms. This can include metadata buried in structured file naming conventions, such as the sample specified by the Inter-Society Digital Cinema Forum (ISDCF) as seen in (Example 1), to structured complimentary metadata embedded in the asset itself or in the form of an accompanying file.



Film Title
14 characters max.
Do not note "3D" unless "3D" is part of film's title. Words should be separated by initial caps: MovieTitle. If film is a sequel with a numeral, add the number without a hyphen: MovieTitle2.

Projector Aspect Ratio
F=Flat
S=Scope
C=Full Container
(See Appendix 7)
If the image is letterboxed or pillarboxed, the optional number after the hyphen reflects the interior aspect ratio.

Language
First 2-3 characters indicate **audio language**. Next 2-3 characters after hyphen indicate **subtitle language**. (XX=No Subtitles)
If present, next characters indicate CCAP and/or OCAP. (See Appendices 1, 1a,1b, and 1c. See Appendix 9 for OCAP and CCAP.)

Studio
2-4 characters
(See Appendix 5.)

Resolution
2K
4K
(Frame rate no longer goes here.)

Facility
3 characters that identify facility. (See Appendix 6.)

Date
YYYY
MMDD

Standard
IOP=Interop
SMPTE=SMPTE
If 3D, add "-3D"

MovieTitle_TLR-1-Temp-RedBand-Chain-3D-4fl-48_F-133_EN-EN-OCAP_US-GB_51-HI-VI-Atmos-Auro-Dbox_2K_ST_Date_Facility_IOP-3D_OV

Content Type:
FTR=Feature
TLR=Trailer
TSR=Teaser
PRO=Promo
RTG=Rating Tag
(RTG-F= Rating Tag for Feature
RTG-T1=Rating Tag for Trailer 1)
POL=Policy Trailer
PSA=Public Service Announcement
ADV=Advertisement
SHR=Short
XSN=Transitional
TST=Test
May include version number after hyphen.

Content Type Modifiers:
If used, use in this order:
Temp=Temp Version (without final picture or sound)
Pre=Pre-release
(Final picture and sound, but without accessibility features.)
RedBand=Special Rating Flag
(Use only for trailers with adult content.)
Chain (Name specific Theatre Chain or Event e.g.: CinemaCon if needed.)
3D=3D
2D=2D version of content that also exists in 3D.
4fl=The mastered luminance if there are multiple versions distributed [e.g.: 35(3.5) 4fl, 6fl, etc.]
48=Frame rate when it is other than 24 (e.g.: 18, 48, 60 etc.)

Audio Type
51=5.1
61=6.1
71=7.1
10=Center Channel Mono
20=LtRt Stereo
If present – and in this order separated by hyphens:
HI=Hearing Impaired
VI=Vision Impaired
Immersive Sound Formats (e.g.: **Atmos, Auro**)
Motion Simulator Formats (e.g.: **D-box**)
(See Appendix 4.)

Package Type
OV=Original Version
VF=Version File (See Appendix 8.)

Territory and Rating
First 2 characters = Territory
Last 3 characters = Rating
(See Appendices 2, 2a, 2b, and 2c.)

Example 1

A common example of complimentary metadata can be found in the Extensible Metadata Platform (XMP) file that is often associated with a JPEG still image. Adobe uses XMP to store metadata through its Creative Suite of tools and many contemporary asset management systems have implemented mechanisms to read, import and associate the data stored in the XMP as part of their environment. Similarly for video assets, wrappers such as MXF or packagers such as IMF can carry significant amounts of metadata for use downstream by asset management solutions or workflow automation devices.

While the metadata carried in these various mechanisms is valuable in the indexing and searching process, the term “garbage in, garbage out” is 100% applicable. If data is not accurately provided through these mechanisms, or if the asset management layer doesn’t have good “ingest hygiene” that includes proper logging and tagging, then the ability to search and index assets against the provided metadata is meaningless.

In addition, the metadata supplied through most of these mechanisms is usually either technical in nature (file format, audio channels, bitrate, etc.) or descriptive at the file level only and not contextually relevant to any time-based metrics. To address this gap we traditionally rely on logging efforts by hand.

Traditional logging is what we have all come to know as eyes and ears scanning a piece of content and identifying visual or aural points of interest based on either a predefined set of discovery metrics, or on the subjective determination of the logger themselves. This method involves a heavy use of manpower and infrastructure, and exposes content to security risks as many times content is shipped out of house to perform the logging tasks in regions with lower labor costs. No matter where it’s done, traditional logging is a time consuming and labor intensive process.

However, even with the best ingest hygiene and sophisticated logging processes in place there will always be search scenarios that cannot be addressed by the breadth of file level data provided, or foreseen given future search demands. As an example you can imagine at the time of a given production there may have been an effort to log scenes where vehicles are present. That same logging request may not have had a requirement to log the types of vehicles in each of those scenes, let alone vehicle makes, models, or colors. At some point in the future for content re-use, there could very well be interest or need to know what makes, models, or colors of vehicles are in certain scenes.

With all of that said, you’ll never be able to predict the future of what metadata is needed or in what context it may be applied. However, by maintaining good ingest hygiene, and through the addition of complimentary technologies that support mathematical indexing of asset essences for future use in dynamic searching (in this case audio and video), then you can have an extensible platform that provides "fused" search results across video, audio, and your organic metadata improving your search results relevancy and accuracy.

2. THE APPROACH

First, let’s get our definition of Metadata Fusion on the table:

metadata fusion - a merging of diverse data sets from distinct or separate elements all describing or providing information about a common temporal referenced point of other data.

To best illustrate the concept of Metadata Fusion we'll rely on a fictional example of a scene from an episodic production called "The Doe's". We're editing a flashback sequence for a new episode and the scene we're hoping to find in our asset collection involves a wide shot of a beach with a white Ford van parked in the foreground and two people walking along the waters edge in the distance. Unfortunately we can't remember what season or from what episode the scene originated. We do know the spoken dialog in the desired scene had something to do with "the problems back in the warehouse".

Now that our test challenge is defined, let's look at the environment we've selected to prove our POC and see how search results relevancy and confidence are greatly improved through metadata fusion.

3. THE ENVIRONMENT

Asset Management Layer (AML)

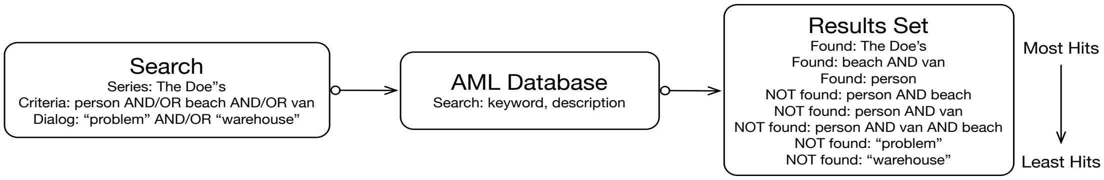
For our POC environment, we have selected a commercially available AML that serves as the aggregation and analysis point for the various metadata indexes we create during our POC. The AML helps facilitate the good ingest hygiene that's required, as well as supporting the basic hand logging processes used to create a base level of organic metadata for each asset. Our POC environment will leverage organic technical metadata such as that discussed earlier and which was acquired during the import process, as well as logged contextual metadata that is entered either at the time of ingest or post-ingest in the AML.

Our selected Asset Management Layer (AML) includes a sophisticated logging infrastructure that reduces the logging burden by providing controlled and automated vocabulary, including "learned vocabulary" whereby the common terms are automatically incorporated into the system and enriched with synonyms and semantic processing. Additionally with the efficient and controlled nature of the AML proxy capability, these sophisticated and highly efficient logging interfaces may be hosted and served to a domain centric, low-cost audience.

While most AMLs provide logging and searching at this level the AML we selected goes a long way to further address the logging challenge by using sematic processing and defining logging terms in the form of keywords with defined ontologies, think of this as a controlled thesaurus with relevance to the theme of the material.

We'll use our AML's search engine as the baseline for all of our POC test searches. The AML also serves as the unified portal (or "lens") for viewing our assets, performing searches, viewing search results and the associated metadata. The AML is also the database-of-record for our proof of concept.

The basic metadata captured during ingest and the logged metadata will contribute the first of the three "diverse sets of data" that will be used for our POC. Searching this associated basic metadata provides an ordered result set of assets that has one or more of the matching search criteria found in the logged metadata associated to the asset (Example 2). A good set of results, but we can improve the relevancy further by adding in our other search engines and correlating the results.


Search
Series: The Doe"s
Criteria: person AND/OR beach AND/OR van
Dialog: "problem" AND/OR "warehouse"
AML Database
Search: keyword, description
Results Set
Found: The Doe's
Found: beach AND van
Found: person
NOT found: person AND beach
NOT found: person AND van
NOT found: person AND van AND beach
NOT found: "problem"
NOT found: "warehouse"
Most Hits
Least Hits

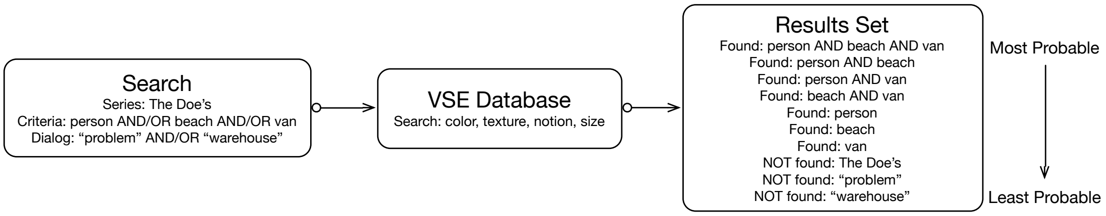
Example 2

Video Search Engine (VSE)

To create the second diverse set of data to be used in our POC, we have selected a commercially available VSE that creates a mathematical model of every frame of video to be used as an index. This allows the system to analyze and index a video asset one time regardless of what the actual search criteria may ask for in terms of an object, pattern, color, texture, etc. Using the created index we can instantiate adhoc searches that address or combine any number of attributes present in the video image. Lets use the example of a "daylight beach scene". We can teach the system what a "daylight beach scene" most commonly looks like (ex: bright blue sky in upper 1/3 of frame, dark blue/white/gray middle 1/3, and tan/white lower 1/3). The VSE can then scan the previously created index looking for frames that exhibit the basic criteria for color, frame location, size, texture, and relationship to other specified search criteria to achieve a usable results set.

Used alone the VSE provides a unique perspective for identifying the desired search criteria. Think about using the VSE to identify specific image details that were not logged as part of the initial logging process, for example a car manufacturer's medallion such as Ford or Chevrolet. You can pull an image from the Internet, or from a defined set of images directly from the AML that you have previously acquired and validated, that looks like the object you wish to find and use it in the system to search for images that match or "look like" the image you provided, or simply mark an area in the actual image essence to be used as the search criteria.

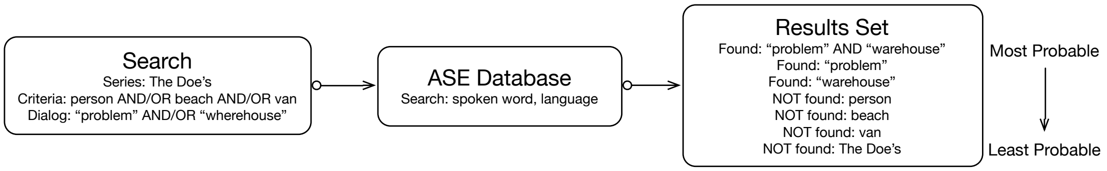
Searching the established index using descriptive definitions or notions, such as what a beach should look like, or adding specific images to the search that represents a person, the system can analyze the image and determine not only the probable match for a beach but also the presence of a figure (person) within that beach scene. The resulting search set returns a higher probably of a match by identifying an asset in which all search criteria are met even though the traditional logging process never tagged that specific asset with those descriptions (Example 3).



Example 3

Audio Search Engine (ASE)

For the final and third leg in our diverse data set triad for this POC, we've selected a commercially available ASE that creates a phonetic index of spoken words for a given asset. In this case we chose a phonetic indexing tool as opposed to a "speech to text" translation tool to avoid the restriction of literal text matching found in many speech to text conversion tools. The phonetic matching ability of the dialogue search tool allows us to do "sounds like" searching for audio elements without concern for spelling and grammatical errors. In addition, the selected tool allows us to search in many non-English spoken languages when needed. The search query returns a data set that is ranked as most probable to least probable based on a search for the provided criteria despite the misspelling of warehouse as "wherehouse" (Example 4).



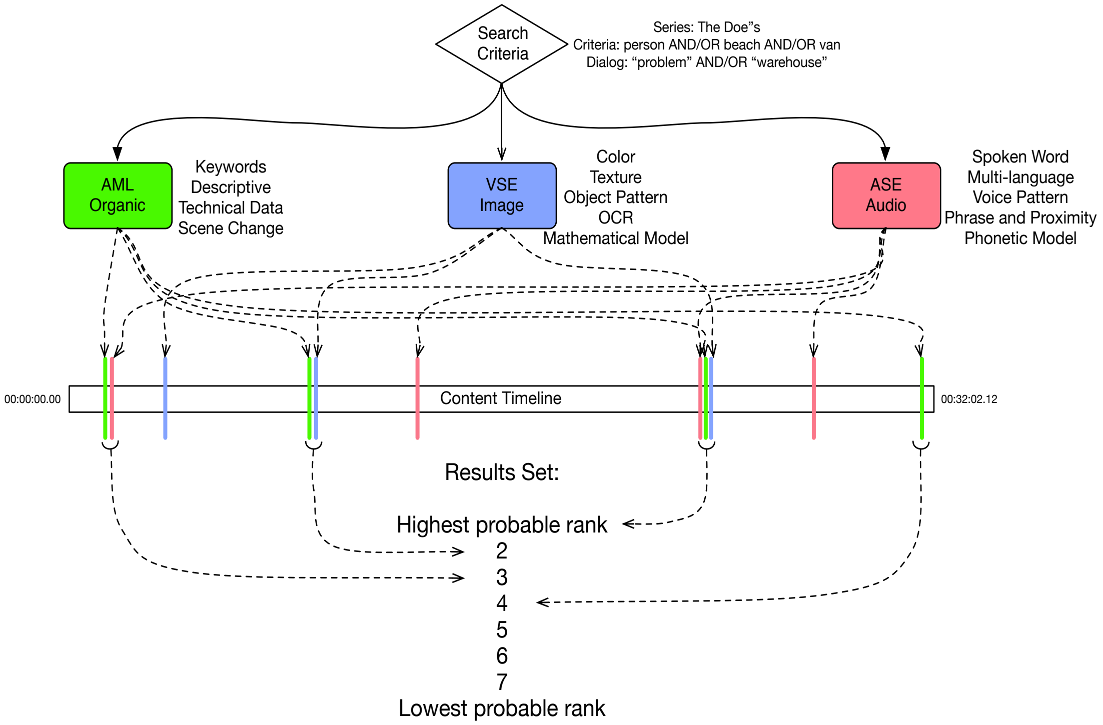
Example 4

To further narrow our search results the ASE can be used to search for a phrase instead of a word. We know John Doe was talking about a "problem at the warehouse" in our fictional show, or perhaps we recall the spot where he was talking about "leaving his van at the warehouse". We can pull a phrase from the script if we happen to have it available as enhanced metadata, or we can build a phrase that can be searched like; "I am going to leave my van at the warehouse", which can increase the probability and confidence of the search results.

4. THE SOLUTION

While each of our engines (AML, VSE, ASE) provides a unique set of search results by searching though a single linear path of organic data, image essence data, or audio essence data, for our solution we will tie these three disparate result sets together using time-based positioning data. Once we have a common time reference we can analyze ranking data from each of the three processes to create a single aggregated and ranked set of results.

The diagram provided in (Example 5) depicts the basic concept for how the three selected search engines will be used to reference the same piece of content on a common time-based scale. Through a logic layer designed to analyse the results and extract highest probability results based on criteria matching, this model spreads the search across multiple essences on the common timeline and quantifies "hits" at any single point in time to each asset essence.

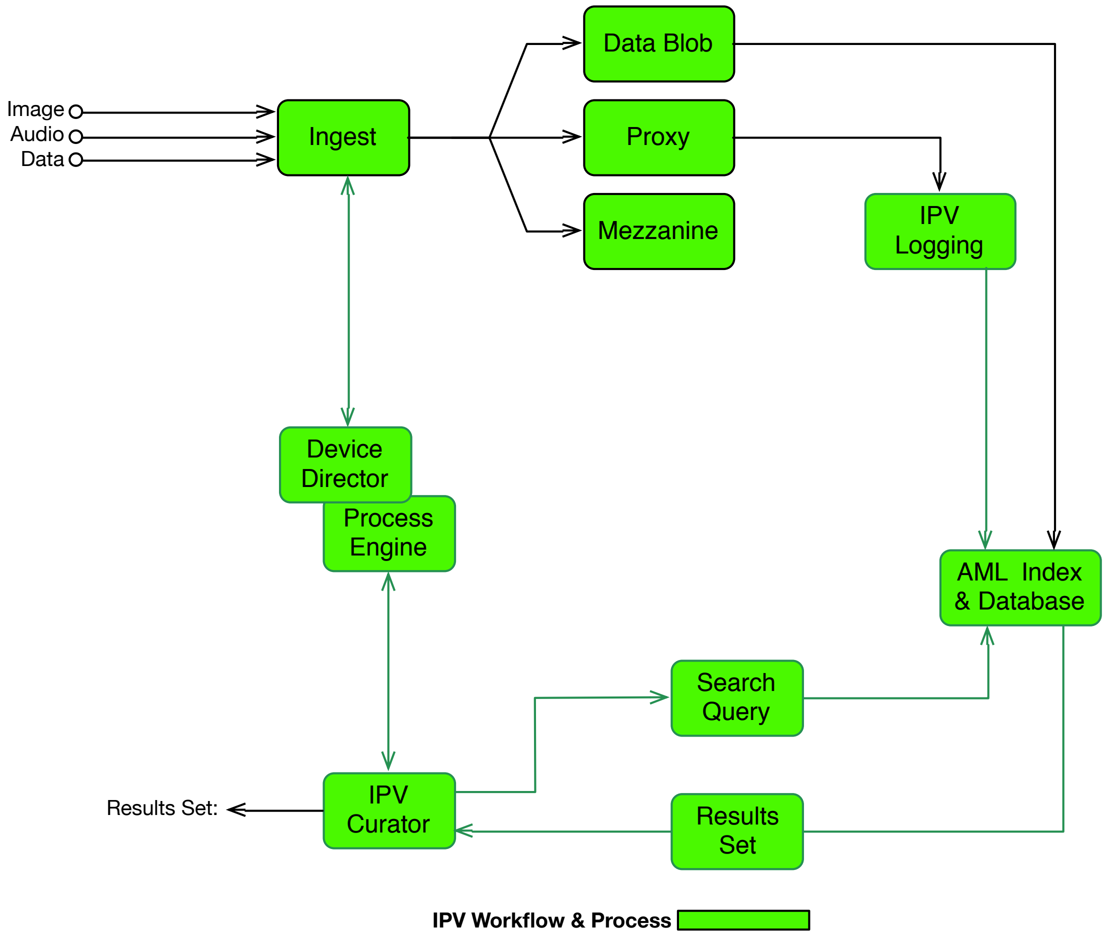


Example 5

This model is extensible and provides a critical framework to allow for future enhancements. The sophisticated algorithms based around the fusion of the results allow for more axes of processing to be added at any time. For example a parallel AML could be leveraged to provide additional refinements to an existing library by taking into account the weighting applied by the independent ranking from the ASE and VSE engines and return that ranking to the primary AML.

The model can also be easily expanded to add other tools to the environment such as geospatial location data, shooting scripts, character data, social media, or 3rd party databases such as IMDB that can further improve breadth and accuracy of the search results.

To power our AML for the POC, StorExcel selected the *IPV Curator* solution (www.ipv.com). This is a complete end-to-end asset management solution for enterprise and workgroup media production and content management. The logging, ingest and searching capabilities of *IPV Curator* enables the efficient population of a media library along with associated metadata to provide our baseline organic metadata, and it will serve as our primary search interface to discover and present our enriched data search results (Example 6). In addition to the *IPV Curator* library management toolset we also leverage the *IPV Device Director* component for device integration and control, and the *IPV Process Engine* component to drive and manage the workflow processes needed to extract, aggregate and associate the extracted metadata from our VSE and ASE tools.

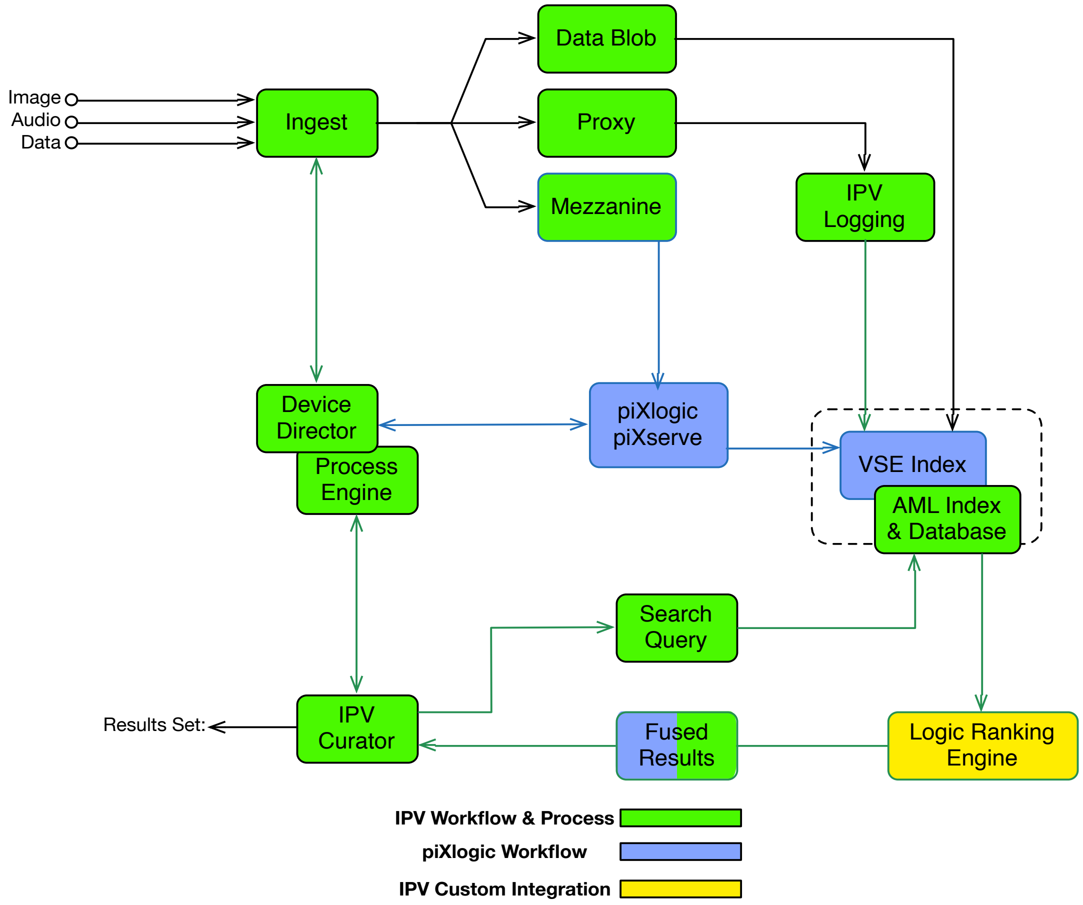


Example 6

The VSE engine that StorExcel chose is the *piXlogic piXserve* solution (www.pixlogic.com). This VSE was chosen for its unique approach to mathematical modelling of the video frames and its ability to automatically index and search image and video repositories, live video feeds, and Internet content. The *piXserve* engine is integrated to the *IPV Process Engine* and *Device Director* through the use of a REST API provided by *piXlogic*.

IPV workflow automation controls the *piXserve* engine as a device to automatically create and capture the image index for future searching at the time of ingest. The *piXserve* engine can be leveraged as a transactional step in a larger workflow such as we did for our POC, or it can be added to an existing library by simply pointing the engine to a repository of video files or live IP-video streams to automatically index their contents. No manual intervention or data entry is required. *piXserve* "sees" what is in the image and automatically creates records that describe the shape, position, size, color, etc. of the discernable objects in the image. In addition piXserve recognizes text that may appear on the image, as well as faces and specific objects of interest to users. In turn these recognition capabilities allow piXserve to automatically create searchable keywords describing the content. In short, whatever is in the image is automatically indexed and available for searching without having to ever re-access the original image.

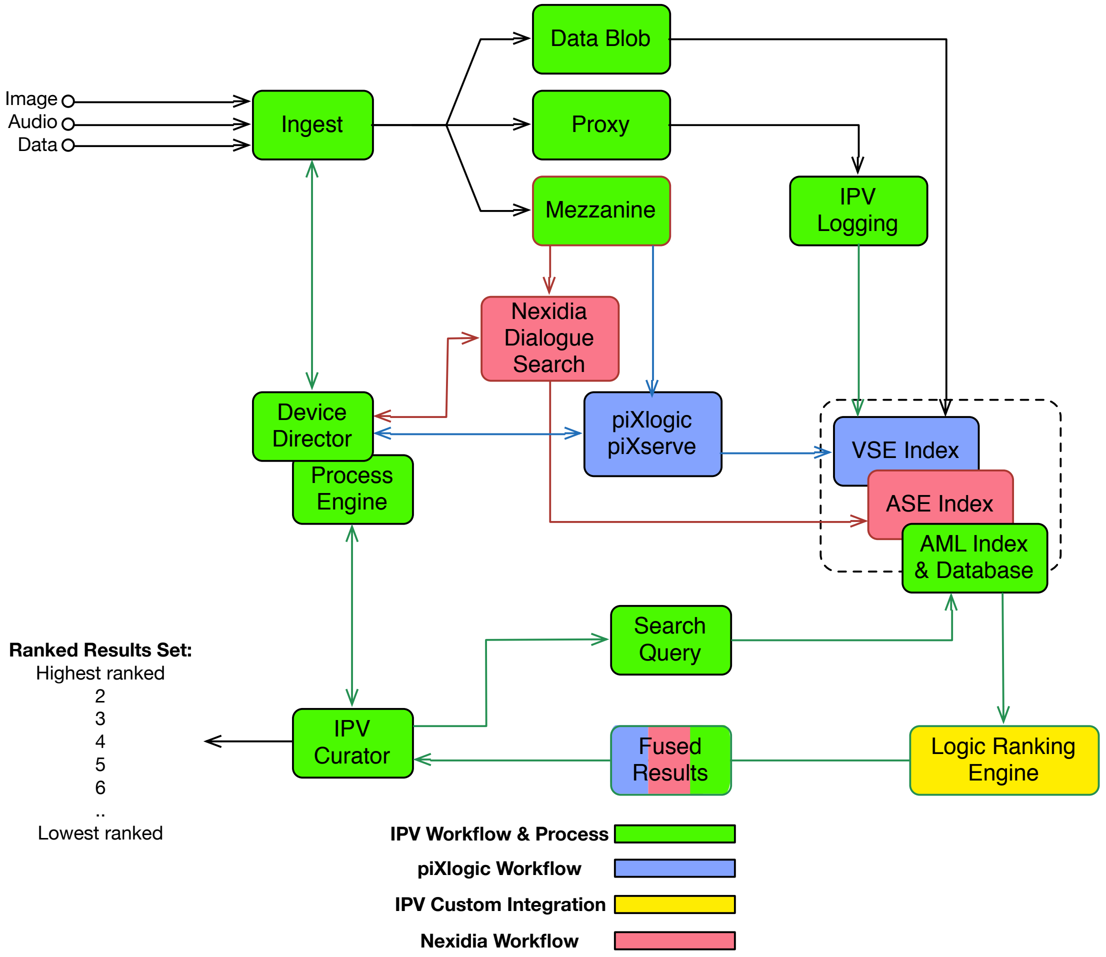
The POC environment now has the ability to query both the VSE Index and our AML database through the APIs that have been integrated by *IPV* and return a raw set of data for analysis. Then through the custom *Logic Ranking Engine* developed by *IPV*, data is correlated against a time domain to create a single temporal index of the query results that can be analysed, ranked, and presented back to the user through the *IPV Curator* interface as relevant fused results set in response to the selected search query (Example 7).



Example 7

Layering in the third leg of our POC is the ASE from Nexidia (www.nexidia.com) that StorExcel chose to enrich the audio dialogue searching capabilities available to the AMLs search engine. Nexidia Dialogue Search technology delivers significant advantages over other solutions. Nexidia's patented phonetic dialogue searching technology quickly finds words or phrases that are spoken in recorded media streams with the ability to search 100,000 hours of media in under a second. Because it analyzes phonemes instead of words, Nexidia's patented technology is far more accurate than standard "speech to text" conversions. It easily recognizes proper names and slang, is not thrown off by individual accents, and does not require perfect spelling or rely on limited dictionaries that need frequent updates. Over 32 different languages and dialects for searching are currently supported and may be configured at the system level for use.

For our POC, the *Nexidia Dialogue Search* engine is integrated into an automated AML controlled ingest workflow through use of the *Nexidia* REST API. The *IPV Process Engine* and *Device Director* controls the ASE in the same way it controls the VSE to create a future searchable index. The raw search results are returned from the ASE and VSE indexes to the AML then passed to the *IPV Logic Ranking Engine* where they are analyzed, ranked and presented back to the user as a fused results set with sophisticated ranking of relevancy. The final ASE and VSE enhanced POC environment is shown below in (Example 8).



Example 8

The POC demonstrates that bringing in additional time-based data can improve search reliability, accuracy, and relevancy. While our POC search example might have been something that could have been logged by hand and discovered by using that same logged data, it would have first needed to have been identified by someone as possible search criteria and logged appropriately. Automating the ingest, logging, and analytic processes to extract all unique metadata not only saves time and ultimately money during the initial processing of the data, but can be instrumental in future needs to meet tight and tough timelines.

Another example

Let’s look at another example that would not have been a scenario that could have been identified in advance. You’ve just found out that your show has been picked up for international distribution in Hungary. Great news! Now you need to prep the content for versioning and there are a few requirements that your distribution agreement has stipulated. One is that all soda cans or bottles must not have visible branding on them, two is that all text within an image must be replaced with Hungarian text, and finally they want to ensure that John Doe’s name is not modified when language versioning occurs so they would like a log of every time his name is spoken and it’s time reference. The good news is that you have all the tools and the indexes already created to make the job fast, less costly, and ensure you meet their delivery requirements.

First they’ve asked for identification of any frames that display a soda can or bottle so those images can be blurred in an editorial session. Our first search is to look for cans and bottles in all episodes and flag the reference frames. A quick web search finds thousands of images of soda cans and bottles from every angle. We’ll download and use a handful of those images as the criteria for our image search through the VSE index. Each frame will automatically be tagged based on their unique time reference with the key word “can” or “bottle” as part of the processing.

They’ve also asked for all English text in found in the image essence of show to be replaced with Hungarian text. Not a problem for our VSE. We can scan through a single episode of the show in seconds and identify common on screen text locations like signage at the warehouse or the text shown on the employee’s badges as an example. Based on search results from the VSE simple process automation can add a tag of “text found” to the original content for each time text is discovered. We can combine our more specific search criteria with some generic images of street signs and billboards pulled from the web and we have a pretty comprehensive set of examples for our search engine to identify and allow replacement of text in an editorial session.

Finally they asked for a time referenced log of everywhere that “John Doe’s” name is mentioned in the audio. For this we’ll create a simple audio search that our ASE can execute across an entire episode or season creating a time reference for every occurrence of the name John Doe, or John, or Doe in any combination. We can collect and retain this data so that post language versioning we can rerun the query and match the findings to our original results validating that the language versioning was done correctly.

Leveraging tools like a Video Search Engine and an Audio Search Engine can greatly improve accuracy of contextual asset information without the need for increased manual labor costs to perform enhanced logging. While results are not yet 100% accurate for either video or audio pattern searching, the technologies continue to improve. Currently we can leverage the use of these technologies to guide labor resources to areas of interest, greatly reducing labor costs and time in the process. Those same technologies can be used in non-critical scenarios today, such as asset library searching, to significantly improve the relevancy of our search results.

5. THE CONCLUSION

Through our POC, StorExcel has demonstrated that commercial off the shelf technologies can be leveraged to add significant value to contextual searching, and automation of content analysis and asset identification, through a single integrated platform by providing a common time reference enabling “point-in-time” correlation of otherwise disparate data points.

Searching each of the various content indexes separately (AML, VSE, & ASE) provides a unique search results set within a narrow range of our POC search criteria. While each results set has viable matches that are presented using this method, none of the individual search processes provide a comprehensive, relevant set of results. By introducing a common time domain and automated processing for correlation of “point-in-time” metadata, we can relate the otherwise disparate results in an intelligent and timely manner greatly improving search accuracy and relevancy of the results through smart ranking. This search capability of providing timely and improved search accuracy is key to future cost effective multi-media library growth and discovery for content monetization.

A key benefit of the use of commercial (3rd party) off the shelf technologies is also to display the modularity of the solution to introduce new components. This can be in exchange or complimentary to the ones showcased in the POC, as well as those already integrated into existing customer environments. With continued expansion of the environment through the addition of tools passing data about geospatial location, script dialog, camera data, character data, social media, EDLs, CDLs, and so on; we can continue to improve the automation and accuracy of asset discovery and broaden the scope of available metadata that is contextually sensitive to our content.

This POC opens the door to explore integration to other existing and future data generating processes that occur throughout the pre-visualization, production, post-production, distribution, and preservation management processes. Metadata and searchable indexes can be automatically collected or created at the point of origination, or at any single point in time, and carried throughout the asset lifecycle without the need to continuously enhance an assets metadata for every new search query that may arise.