

# USING AUDIOVISUAL/AUDIO ARCHIVES FOR AI TRAINING

AI Technology Companies (“AI Tech Co”) increasingly seek what archives hold—rich media collections to train their multimodal and large language models (“AI”). Many archives have already had their online collections scraped without permission and are now facing offers from companies eager to secure further access.

To support archives navigating these pressures, we are developing decision-making rubrics to help institutions assess collaboration opportunities with AI Tech Cos. These tools are designed to guide archives in weighing whether such deals represent sustainable business opportunities that strengthen their futures—or risky bargains that could compromise their long-term best interests. This paper serves as part of that tool and is intended to help as a comprehensive resource for guidance on some key considerations and critical questions to ask.

In this paper we will cover:

- [What is Generative AI?](#)
- [What type of companies want to obtain your data?](#)
- [10 Key Things to Consider when looking at AI Tech Companies and why this is important](#)
- [Five more key areas to think about \(introspective\)](#)
- [Your Internal Technical Preparation](#)
- [Questions to ask the AI Tech Co.](#)

## WHAT IS GENERATIVE AI?

- Generative AI (“GenAI”) refers to the use of AI to create new content, like text, images, music, audio and video.
- GenAI models are trained on very large datasets from which they learn the patterns and structure and then generate new synthetic content that has similar characteristics.
- More and more content globally is being created using GenAI tools, and audiovisual archives could be used to train data sets, should their rights and permissions allow it...
- However, accuracy, authenticity and “truth” is a growing matter of concern
- And GenAI content brings about new legislation and copyright/ownership questions
- Audiovisual archives are challenged to implement policies to address these points; but the archive does not work in isolation from its wider organisation and must collaborate with stakeholders to design its policies

## WHO WANTS ARCHIVE CONTENT?

Types of AI Tech Companies:

- **'Informed' aggregators:** Companies knowledgeable about the space with industry contacts who connect content libraries and negotiate deals.
- **'Digital Locker' businesses:** Organisations collecting content to become intermediaries between content owners and AI Tech Cos / LLMs, handling content delivery for a percentage fee.
- **'Jack of all trades' aggregators:** Companies seeking to acquire multiple rights types (FAST, VOD, AI) with limited transparency, typically offering 50/50 revenue sharing arrangements.
- **Bespoke language model builders:** A fourth emerging model where businesses build custom language models based on specific content and license these models to various clients, creating subscription-based revenue streams.

## 10 HEADLINE CONSIDERATIONS FOR AV ARCHIVES CONSIDERING ALLOWING CONTENT USE IN AI TRAINING

- **Intellectual Property Rights:** Ensure clear agreements on ownership and usage rights of AI-trained models
- **Content Licensing:** Develop specific licensing terms for AI training purposes
- **Data Privacy:** Implement measures to protect sensitive information in the content
- **Quality Control:** Maintain oversight on how the content is used and represented in AI outputs
- **Ethical Use:** Set guidelines to prevent misuse or creation of harmful content
- **Brand Protection:** Ensure AI usage doesn't negatively impact the archive's reputation, or 'brand'
- **Legal Compliance:** Stay updated on evolving laws regarding AI and copyright, *which may differ across jurisdictions/territories*
- **Content Valuation:** Assess the long-term value of your content in the context of AI training
- **Technology Partnership:** Carry out thorough due diligence, how credible are they and are they in this for the long-term, do they have a good track record of working with others, do they have robust security and ethical standards
- **Revenue Sharing:** Consider models for profit sharing from AI-generated content and work with AI Tech Co's who will ensure you really benefit from a financial return.

## WHY IS THIS IMPORTANT?

**Before moving forward, consider philosophically why are you doing this? In what way does this promote the archives values?**

Archives should carefully weigh these factors to protect their interests while potentially benefiting from the emerging AI landscape. The use of audiovisual and audio archives for AI algorithm training is a rapidly evolving field with significant promise for knowledge discovery, preservation, and access. However, it is equally fraught with legal, ethical, technical, and cultural complexities. Success depends on a commitment to robust governance, transparency, documentation, and human oversight, as well as ongoing dialogue among archivists, technologists, legal experts, and affected communities. Only through such a multidisciplinary and principled approach can the transformative potential of AI be harnessed while safeguarding the integrity, rights, and diversity embodied in audiovisual heritage.

## FIVE MORE KEY AREAS THAT NEED FURTHER CONSIDERATION BEFORE CHOOSING TO WORK WITH AI TECH COMPANIES

Let's take a dive into five key areas for further consideration:

1. Strategic, Reputation and Risks
2. Rights and Permissions Pertaining to Your Data/Media
3. Rewards
4. Resource Allocation and Potential Cost Impacts
5. Technology Considerations

### 1. STRATEGIC, REPUTATION AND RISKS

- Why are you doing this—what are the benefits of allowing access to your content?
- Are there risks associated with allowing access to your content; does it devalue your content in any way ; are there consequences for your business model?
- Have you vetted the AI Tech Co and compared them with others?
- Consider what is valuable to your business versus what is valuable to the AI Tech Co?
- To what extent can you define the deal terms on offer, or is the AI Tech Co fixed in their offer without space for negotiation?
- Are you granting access to all or some of your archive? Can your content be segmented to meet an AI Tech Co's specific needs and interests?
- If there is a breach/technical problem with an AI company, directly or indirectly impacting your media, how might this impact your reputation and relationships with your underlying rights holders? If this problem puts the AI Tech Co out of business, how will you get your media back?

- Some of the AI Solutions can generate harmful content of the persons in the media being shared with the AI Tech Co, for example Vishing Scams. Because of the widespread abuse already to distort various types of media, audio, images and/or video considerations must be taken in this case. Have you considered this possibility?
- With the level of risk for dealing with an AI Tech Co, careful consideration as to what AI Tech Co is chosen must be made. If you go with an unknown, small and new business there is no way to guarantee you will get your data back, or secured if that company is gone tomorrow. This can be another way of your data/media getting in the wrong hands. If the AI Tech Co goes out of business suddenly your contract/agreement with them is no longer viable. Consider only looking at larger reputable organizations?

## 2. RIGHTS AND PERMISSIONS PERTAINING TO YOUR DATA/MEDIA

- How is your data/media being used by the AI Tech Co?
- What is the right that you are granting to them - how is it defined - and can you grant that right?
- Does your content have ethical, compliance or data privacy considerations?
- Does your content have talent (actors, writers, presenters, contributors) attached and how is this handled in any deals?
- Does your content have audio/music attached and how is this handled in any deals?
- Who owns the resulting AI outputs if they are based on your content and data? If your data is being used to train the AI Engine, discuss with your internal groups that the trained media/or output might not be owned by your org.
- At the end of the agreement, what happens to the content/data that you have supplied? What protocols are required for data retrieval or deletion should the partnership end?
- What controls does the AI Tech Co have in place to ensure that they adhere to the grant of rights and that your content/data isn't misused.

- What is the established procedure for addressing compliance concerns with content after machine learning implementation? Are there any penalties associated with non-compliance?
- Any engagement with an AI Tech Co that puts existing licensing, agreements or copyrights at risk from misuse can potentially put your organization at risk legally, however if your data is already publicly available on the open internet it will be very difficult to track who misused the data.

### 3. REWARDS

- What is the AI Tech Co's compensation model - will you be paid by volume of content, type of content, etc.?
- How long will you have to wait in your agreement to see a payment for sharing your content?
- If your content is being used to train the AI, consider the potential for low profit margin or any profit at all since the output could potentially not be owned by your org.
- Will you receive payment from the AI Tech Co purely for use of your data? Or will you have to wait until they make a profit? How transparent are they willing to be with this?
- Does certain content (e.g. by genre) attract a premium?
- If you have data to accompany your content, does this attract a premium? (data could be: descriptive, technical, rights, offline e.g. papers)
- If you have "offline content" (rushes, production data, etc) that can only be accessed by going through you, does this attract a premium?
- Are the upsides purely financial or can you get anything more from the technology relationships e.g. support (incl finance) to unlock the required content, data enrichment, catalogue enhancements?

- If your content is being used to develop AI tools, would your organisation benefit from use of the improved tools and can your org have access to those tools - and include this in the deal terms?
- Consider whether you can obtain enhanced data derived from content you provided to the deal for your own use. Establish clear terms for accessing, using, and redistributing these enhanced datasets, including specific exclusivity rights and usage limitations.

#### 4. RESOURCE ALLOCATION AND POTENTIAL COST IMPACTS

- At the earliest possible stage (before entering into a deal), you will need to assess if you have what an AI Tech Co is potentially looking for (volumes, subject matter, tech spec, etc.) - who will do this work?
- Will you have any resources available to work alongside the AI Tech Co to ensure the correct data is being shared and operational processes are acceptable to the original agreement?
- Have the resources that will work alongside the AI Tech Co provided an internal impact assessment on what they think the actual impacts would be for the work?
- What effort is needed to make an actual deal work? Who are your internal (or external) stakeholders, are they all involved and do they have the time to input to the decision making process?
- Will you incur any costs to enter into a deal e.g. legal, rights, technical or operational resources? This can include legal oversight, technical assistance in confirming your organisation is prepared to interact with the AI Tech Co, and potential technical augmentation to prepare and secure your internal technical systems.
- Who in your organisation would own and approve a deal with an AI Tech Co?
- Do you need to create or prepare assets and data in order to deliver to the AI Tech Co and, if there are costs to make your content ready, will the AI Tech Co contribute towards helping to unlock the content and make it ready? (see further Technology Considerations below)

- If you are considering sharing your data, can you include considerations for watermarking your data for that particular sharing? This will keep your data traceable and ensure if data is leaked, you can trace it to a particular vendor. Consider this extra cost as added insurance for your data. (however, if some of your data is already publicly available it will be difficult to trace that data back to the contracted AI Tech Co.)
- Are you incurring one-time costs, or will you have costs and resource allocation ongoing throughout the life of the deal?

## 5. TECHNOLOGY CONSIDERATIONS

The detailed technical questions included are intended to support a serious look inward at your organisation's technical capabilities, readiness and intentions so that you make careful decisions when planning to work with an AI Tech Co. Before you move a conversation forward with outside AI vendors, seriously consider why this is something you want to pursue? This is not something to rush into, there is no guaranteed financial return. There is a much higher risk of financial loss, reputation damage, and technical issues if the correct decisions are **not** made. As many institutions would like to entrust all the technical work (connectivity, moving the data assets, etc) to the AI Tech Co and leave all of the heavy lifting to them, there is even greater risk in allowing the decisions to happen outside of your organisation, by that AI Tech Co. Do you really want to be the first in line for AI capabilities, even if the full business benefits are not yet defined? In a rush to get something new and potentially valuable (or perceived as valuable) for your organisation, not vetting a company enough or asking yourselves the right questions could take your organisation into a very negative situation.

## INTERNAL TECHNICAL QUESTIONS

### THINGS TO ASK YOURSELF PRIOR TO DEALING WITH AI TECH CO (OR ANY TECHNICAL COMPANY)

The detailed technical questions included are intended to support a serious look inward at your organisation's technical capabilities, readiness and intentions so that you make careful decisions when planning to work with an AI Tech Co.

- How up to date is your technology? Are you aware if the technology that houses your data/media is connected to any other sensitive organisational data like client information or employee data? Understand your technical capabilities. Your company Roadmap should include internal technical preparation prior to engaging your media with an AI Tech Co.
- Do you manage your own technology around your data/media? If not, and the technology that houses your media is managed by an outside party, have you shared your intent to interact with an AI Tech Co with that outside Media Management party?
- What questions have they (the outside company that manages your data/media) provided to you about this? Do they have restrictions? Have they suggested specific technology design specifications that the AI Tech Co would have to adhere to?
- Will your outside technology partner that manages your media today have any restrictions or requirements that will incur costs for potential changes? Will they require a new Managed Services Agreement?
- If you've identified the data/media you want to use with the AI Tech Co, is that data/media secure and separate from your other technology or at least isolated from data/media you do not intend to use?
- Can you include your technical staff or someone that manages your data/media that can speak to the AI Tech Co to explain your systems and your technical requirements?

- If you are not aware of your technology limitations and whether your media is properly segmented, you would need to identify persons (internal or external consultants) to do this. What is the timeline for this review? Based on this review, what is the timeline and cost for preparation with a potential AI Tech Co?
- Have you researched AI Tech Co's to interact with? Have any AI Tech Co's approached you? Have you reached out to any other organisations that may have worked with those AI Tech Co's? What was their experience? [This also plays into the AI Tech Co's credibility]
- What function is the AI Tech Co providing to your organisation? Do you really require AI to make these changes for you? If it's not about training for the AI application and maybe you are looking at services for Up-resing or resolution refinement or media restoration? Do you already have an existing application that might better suit your organisation until you can be assured that your technical software and hardware is secure and ready to connect with the AI Tech Co for services?
- Have you priced similar tools that you might be able to purchase and bring into your organisation (and get training on) instead of risking your data being shared with an outside AI Tech Co? (For example, OCI Media Flow by Oracle).
- Have you talked to Technical Security specialists to discuss best practices to help you prepare your data and systems for dealing with an outside AI Tech Co?
- Have you researched the potential risks with allowing the AI Tech Co to train their AI models on your data? For example: allowing an AI Tech Co to use your media for its training runs a risk of that AI application creating derivative works that look very similar to the media you provided which could lead to larger problems with licensing and even potential devaluation of the original media. (Also can lead to harming your media- See Risks section above)

## QUESTIONS TO ASK THE AI TECH CO

Something to think about when shopping around for AI Tech Co's is that everything is a sale for them. You are dealing with sales people first. They will tell you 'yes' to anything to make you comfortable about choosing them and allowing them access to your data. So although they might not be technically capable of doing some of the technical changes you inquire about, they will tell you whatever is needed to make you comfortable about doing business with them. Ask them pointed questions, if you are not technical, have your technical staff or consultant on hand and ask them direct questions. Do your homework about that company. Anything that you want to enforce about your concerns and use of your data you should be able to get in writing in an agreement with that AI Technical Company. Something else to think about: Some of the largest data breaches and exposures of internal proprietary data were done because of incorrect configurations with an AI Tool/Application (See McDonald's/Paradox; ChatGPT Conversation History Leak; T-Mobile API Misconfiguration; etc).

One additional note: If you are considering discussion with an AI Tech Co or any Technical Company for work on your archives for support for things like up-resing or anything other than use for training the AI Tech Co's AI Engine, many of these questions will help you plan your goals as well.

### CAPABILITY

- Can they assure their application is not using copyrighted material? (where applicable) Or is it at least annotated? How do they do this?
- Have they done PoC's (proof of concept) with other organisations? Are they willing to share any of the inputs and outputs from those PoC's? Share results.
- Are they able to do a demonstration of their product for you without using your data/media?[This also plays into credibility]
- If another company manages your data/media for you, share the restrictions with the AI Tech Co - and do they have any issues with these requirements?

- How exactly will the AI Tech Co's product work? Will they connect to your internal systems?
- How exactly do they propose grabbing your approved data/media and importing it into their AI application? In other words, how will they physically connect their systems to yours to obtain your data?
- Can they provide a visual design showing a proposal of their systems connecting to your system and explain what will be used?
- How will they process your data? Can they share their design for this or is the actual AI processing a black box?
- Will the AI Tech Co allow you to review and approve the system design they propose on how they plan to connect to your systems (in order to grab the data/media) before implementing it?
- Can the AI Tech Co ensure that they provide a Secure Enclave for the data that cannot be co-mingled or shared with the AI Tech Co's other client's data?

## PERFORMANCE/ACCURACY

- How do they (the AI Tech Co) measure and evaluate their application's performance? How do they monitor this?
- For Technical Services like Up-Resizing: How do they handle hallucinations? What does their mitigation plan look like for corrections? What sort of time frame could that be for corrections (potentially)?
  - For example: super refined resolution can add objects that weren't there before or create unnatural/incorrect movements. Phantom voices, incorrect words or changing speaker identities.
- For Technical Services other than AI Training: How do they separate their specific services like image repair or Up-resing from the AI Training?

- Does the AI Tech Co also provide data repair separate from the AI Training, and can they explain how they keep the data being repaired separate from training data?
- Can the AI Tech Co provide performance metrics or error rates?
  - For example: For audio restoration or refinement - do they have a word error rate?
- How do they, (AI Tech Co) handle scalability?
- Does the AI Tech Co allow you to review and approve the final output of your data/media if you do use their AI services?

## PRIVACY/RISK/SECURITY - AND MORE ABOUT CREDIBILITY

- Request a Security Assessment/risk report from the AI Tech Co.
  - Why this matters: if the AI Tech Co is directly connecting to your organisation's infrastructure/systems to pull/push data/media for this AI process, knowing if they have had an assessment, or if there is a risk report on their systems can help with decisions on moving forward. This would be good to understand if they are going to integrate with your systems to intake your data/media. Also to help in understanding the AI Tech Company's risk factor of doing business with them. Risk = Likelihood x Impact.
- Has the AI Tech Co indemnified themselves from hallucinations?
  - This question is handy if you are using the AI Tech Co for professional services like "Up-resing".
- Does the AI solution/design from the AI Tech Co rely on Third party vendors or other outside dependencies? If so, how are these dependencies secured? Is it possible for them to do a Software Composition Analysis check?
  - A Software Composition Analysis (SCA) check is an automated process done by a trusted outside party (cyber security firm) and checks the system

components for vulnerabilities and creates a report outlining this called a SBOM, or Software Bill of Materials.

- What best practices are followed by the AI Tech Co to monitor and secure the AI chain mitigating risks with other 3rd party vendor components?
- Is there a fence between your organisation's assets (data/media) and other client's assets being stored and processed by the AI Tech Co within their systems?
- How does the AI Tech Co restrict the use of your data/media exclusively to purposes you've authorised?
- How does the AI Tech Co ensure that your organisation retains full ownership of all data/media put into their AI system(s)?
- What assurances can they/AI Tech Co provide that your data/media will not be used for training, marketing or analytics without explicit consent?
  - This question is applicable in various forms regardless of how you engage with the AI Tech Co. Maybe you will allow them to train their AI Engine but you don't want your data to be used in demos/advertising.
- Is the development team of the AI Tech Co off-shore? How will they ensure that all data/media, this includes inputs and outputs and training data (where applicable) remains within the geographical boundaries specified by your organisation?
  - For example: If the AI Tech Co is within your country's location and adheres to your data/media privacy laws. Their AI development teams might not be within the same geographical boundaries. If your organisation is located within the United Kingdom and the AI Tech Co is as well, however their AI development teams are located in Belarus, or India, or another country not bound by your location privacy laws. How could you enforce the security of your data? Ask the AI Tech Co not to allow offshore development or storage with your original data/media.
- Is their AI Engine/process within a secured closed loop and not touching the open internet?
  - This is not about user access (user access is outlined in the bullet directly above). This is about the AI Tech Co's actual AI Engine/Application and its

overall design. If their design is exposed to the open internet, your data/media will not and cannot be secured regardless of their assurance, and is very much at risk.

- If the AI Tech Co is directly integrating their systems to your organisation's systems, ask them if they would be willing to do an AI specific pentest ("Penetration Test") on the systems once integration is done?
  - A Pen Test or Penetration Test is an authorized test/attack done by a security personnel or security testing person to check system vulnerabilities in the network, application, or systems. A simulated cyber attack that helps identify areas of concern so that they can be corrected to guard against bad actors.
  - Pentesting could be an added expense but necessary to ensure systems and their configurations are as secure as possible. Also find out who will be responsible for the Pentesting costs since it is not uncommon for companies to use outside consulting parties for this.
- Has the AI Tech Co had a Pentest done on their AI model before?

## POTENTIAL COST/PRICING MODEL

If you are shopping for AI Technical Services (restoration, Up-resing, etc) instead of being approached for AI Training, aside from the hidden costs listed above, below is a very high-level guide for potential cost models.

- Based on your organisation's requirements, what does their cost/pricing model look like? Do they have tiered pricing?
- If you allow them to train their AI Engine with your data, do you get reduced pricing for this other service?
- What is their cost based on? Is it per image/or track? By transaction? By minute or compute hour?

- What is their timeline for delivery of your requested service? What sort of discounts do you get if the product isn't on time or not to your specifications (and needs corrections)?
- If you've checked the AI Tech Co's references, if they used them for services other than AI training, did they feel comfortable with the costs of the service? Were there any surprise/hidden costs?