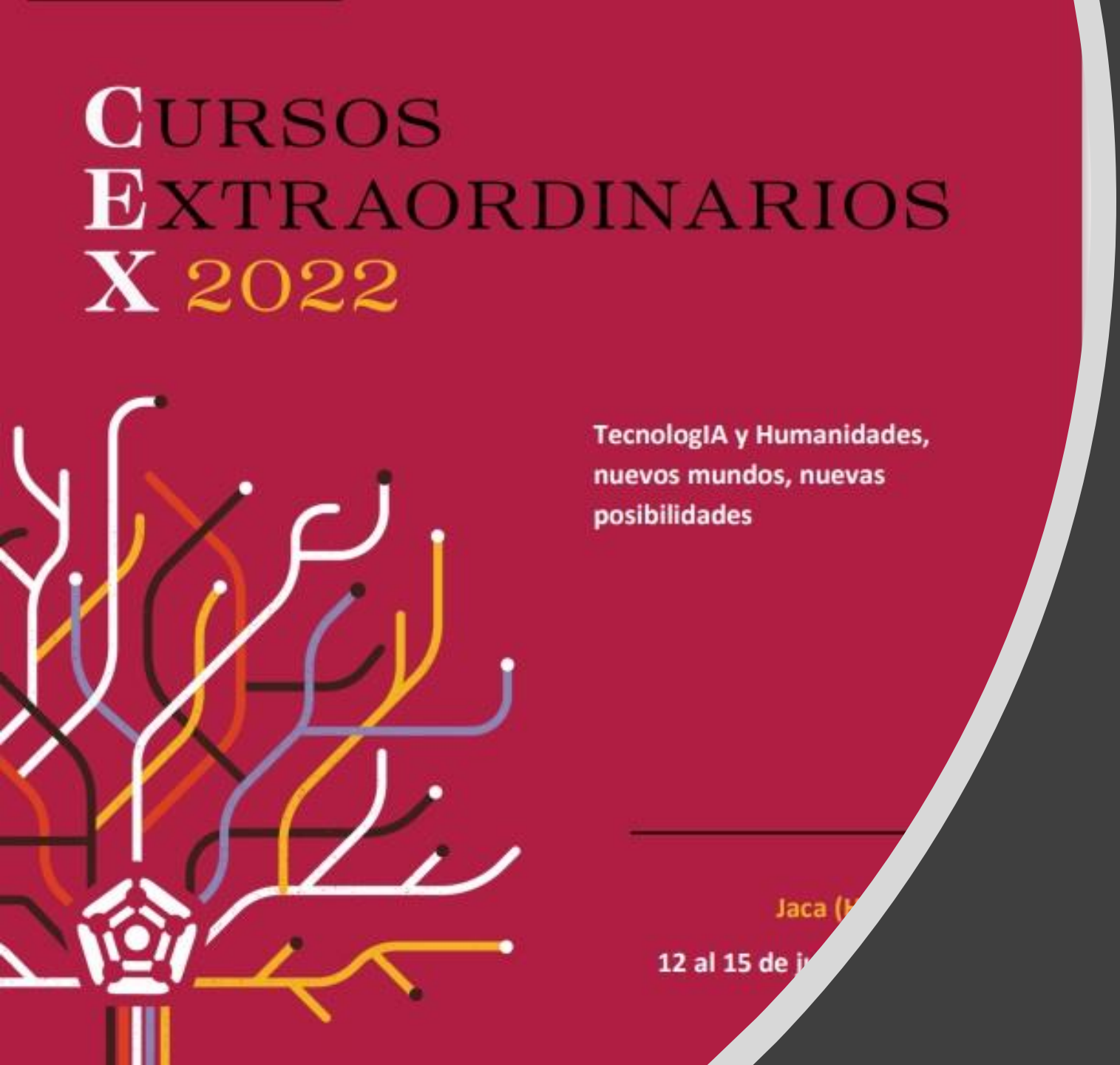


Metadatado automático en el Archivo de RTVE

Carmen Pérez Cernunda y Virginia Bazán-Gil

RTVE

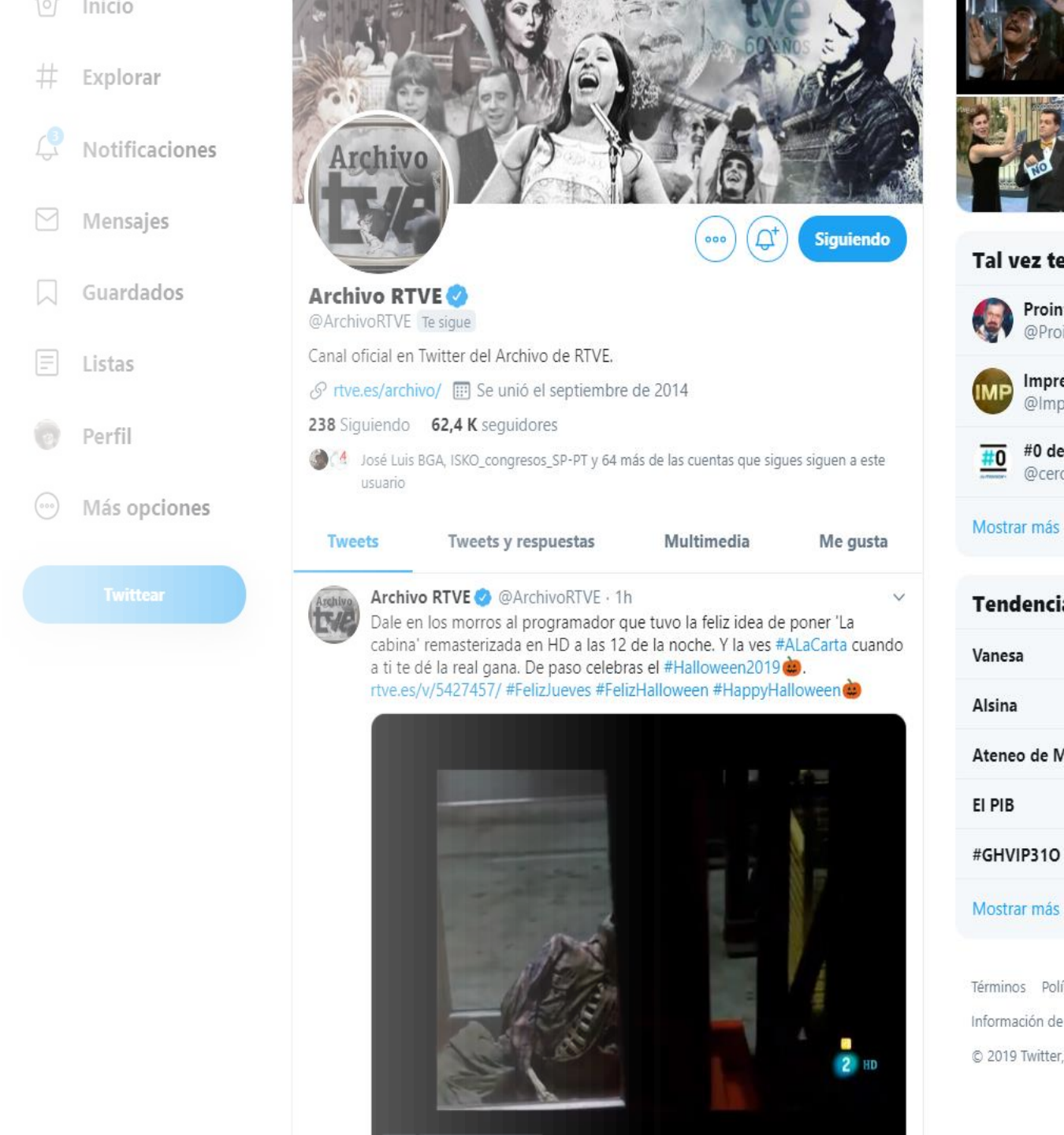
6. Mejor proyecto de innovación

Ganador: RTVE
Aplicación de IA y cloud para el análisis de contenidos en el archivo de RTVE

En el marco de la digitalización y en consideración de su entidad de servicio público, RTVE hace ya algunos años llevó a cabo la transformación a archivos digitales de todos sus fondos documentales. Este proceso ha hecho que los usuarios accedan a más de 2 millones de horas de contenido. Sin embargo, durante este proceso de transformación hubo una parte de la media digitalizada cuyo contenido no pudo recibir ningún tipo de tratamiento documental, lo que significa que, aunque estén disponibles, son fondos que difícilmente se pueden llegar a utilizar ya que los usuarios no tienen conocimiento

Misión del Fondo Documental RTVE

- Preservar los archivos de RTVE
- Garantizar la accesibilidad del patrimonio audiovisual de la corporación
- Facilitar la puesta en valor de los contenidos para la producción y la comercialización.



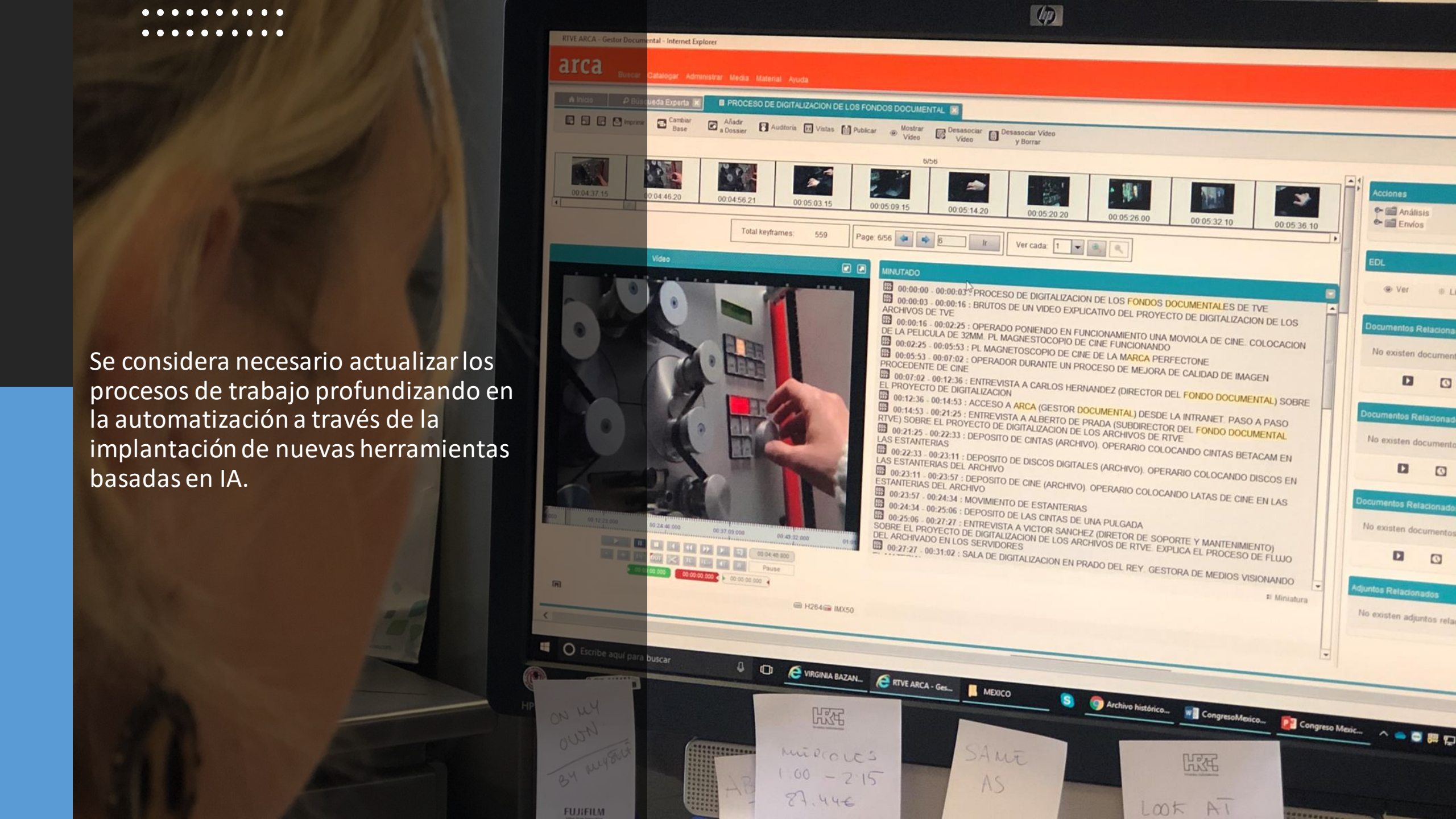
En la última década

La digitalización de los procesos productivos en RTVE y la digitalización de los archivos de RNE y TVE ha permitido

Un fácil acceso de los usuarios a los fondos de forma rápida y sencilla

El incremento en el uso para la producción de nuevos contenidos, la reemisión, comercialización y difusión a través de nuevas plataformas.

Se considera necesario actualizar los procesos de trabajo profundizando en la automatización a través de la implantación de nuevas herramientas basadas en IA.



Dirección de Estrategia Tecnológica

Fondo Documental RTVE

Cátedra Universidad d e Zaragoza

Objetivos del proyecto

- Análisis automático de más de 11.000 horas de video en 1 año.

- Preparar ARCA 3.0 para recibir metadatos externos, independientemente del proveedor de IA.

Funcionalidades requeridas: audio

Transcripción de audios a texto, debidamente puntuado, capitalizado y con reconocimiento de hablantes.

Reconocimiento de entidades nombradas o extracción de entidades: identificar y clasificar en categorías predefinidas como personas, organizaciones, lugares, fechas, etc.

Extracción de palabras clave

Clasificación automática de contenido

Extracción de caracteres alfanuméricos en imagen, en forma de rótulos y/o subtítulos.

Reconocimiento facial y de identidad

Identificación de objetos y reconocimiento de escenas.

Reconocimiento de logos y marcas

Funcionalidades requeridas: video

Cronología del proyecto

Duración Proyecto:

- 4 meses configuración y desarrollo (Fase I)
- 12 meses de servicio (Fase II) prorrogable un año más

Inicio contrato. Fase I:

- 1 Mayo de 2021 (Fase I)
- Se acordó un mes más para configuración

Inicio Servicio. Fase II:

- 1 octubre 21 – finaliza 30 de septiembre 2022
- Aprobada prórroga hasta septiembre 23

Áreas implicadas en el proyecto

rtve

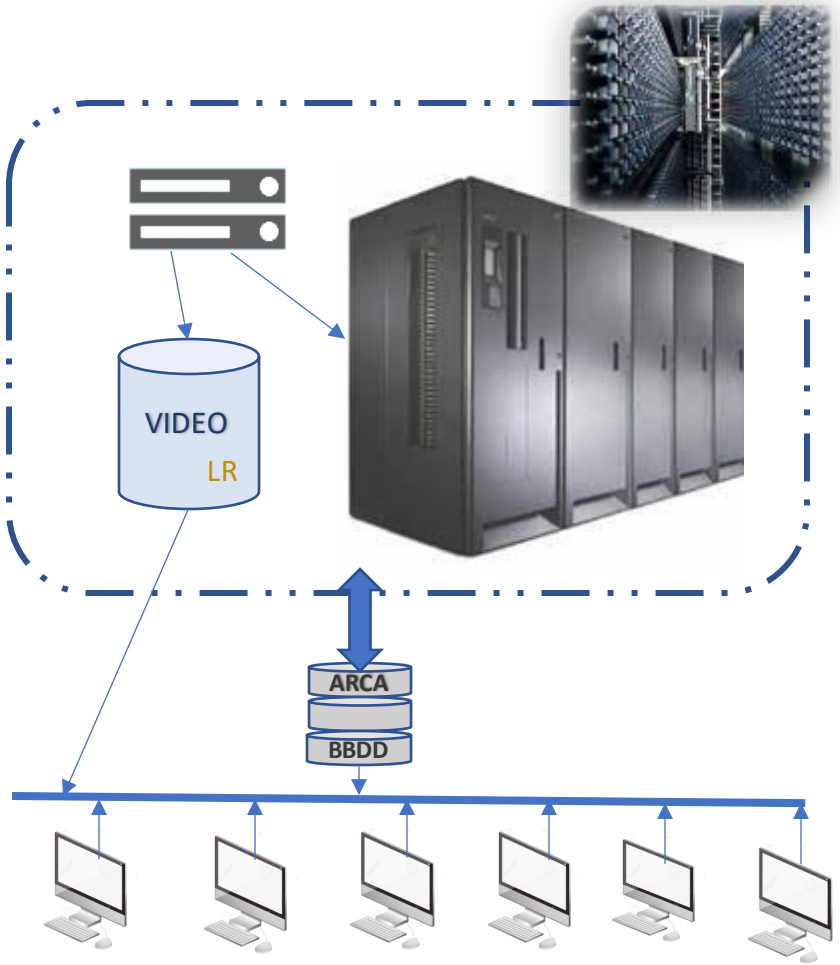
- **Innovación**
- **Fondo Documental**
- **Sistemas**
  - **Desarrollo de Aplicaciones**
  - **Comunicaciones**

VSN
INNOVATION & MEDIA SOLUTIONS

ETIQMEDIA

Archivo RTVE

¿Cómo es?

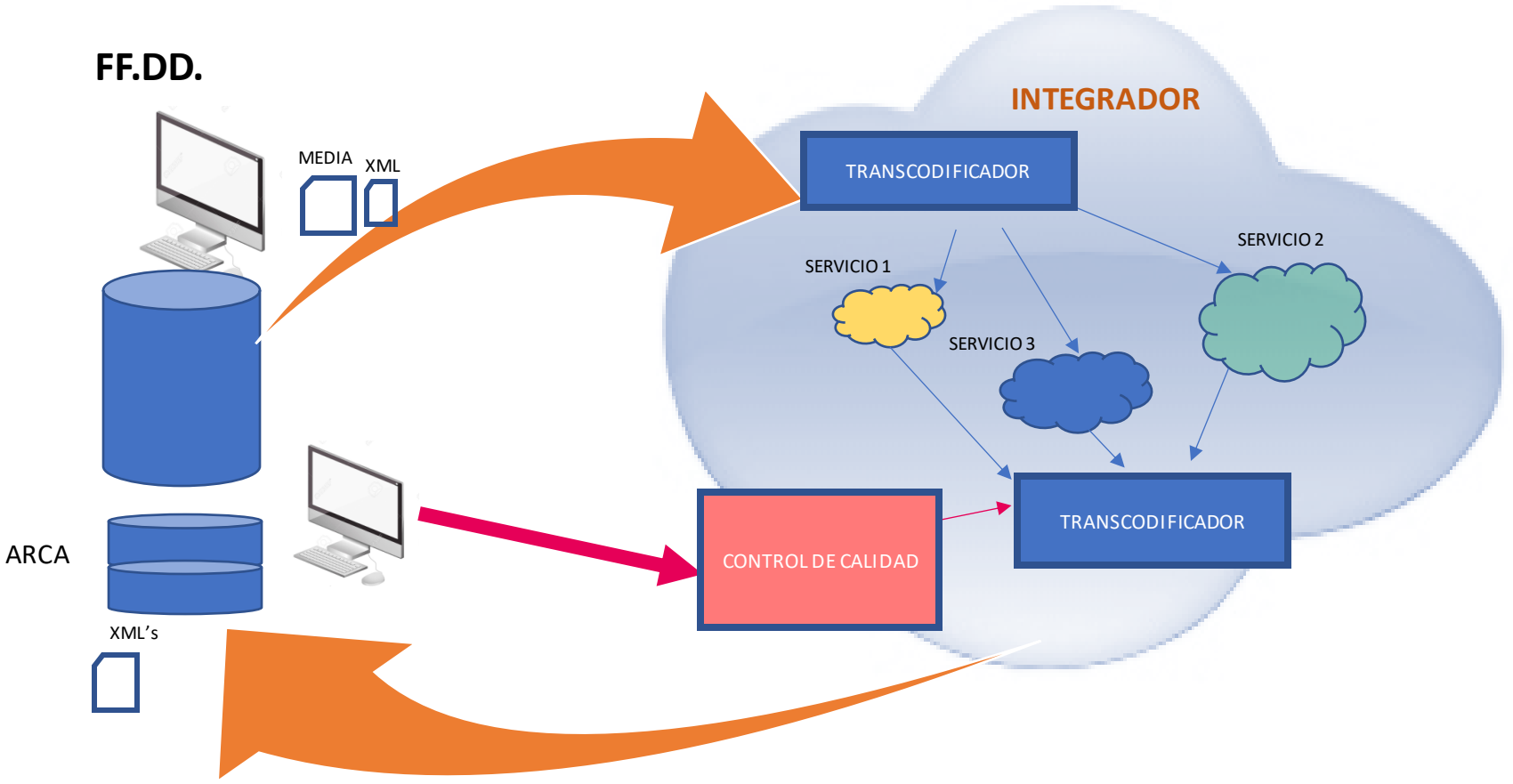


- Librería
- Sistema discos duros
- Sistema de control
- Gestor Documental: ARCA
  - ✓ La herramienta usada por documentalistas
  - ✓ BBDD con metadatos asociados a la media

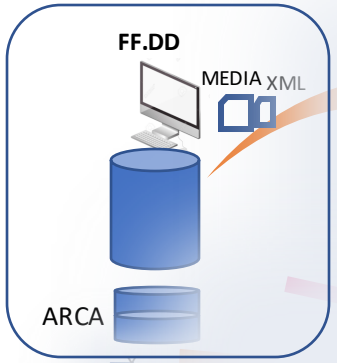
Ficheros LR: H264 a 1,5Mbps
hasta 8 canales de audio

Expediente Metadatado automático

ARQUITECTURA



Inicio del proceso: en RTVE



VIDEO
LR

Selección Media → Descarga en FTP → Envío a VSN

	Selección Media	Descarga en FTP	Envío a VSN
Quien ?	Documentación	Automatizado	Automatizado
Herramienta ?	ARCA	ARCA	Automatizado

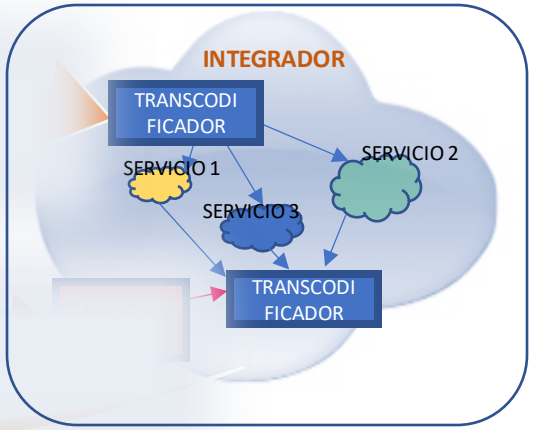
Puede ser uno o varios archivos
Permite selección de tratamiento: solo audio o audio+vídeo

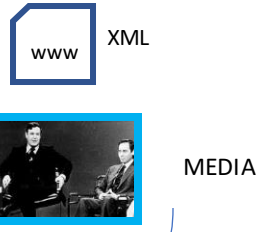
Para cada contenido: Fichero media + XML con información sobre:

- Tipo de Media
- Audio…
- Tratamiento seleccionado

Cada hora se envía todo lo que ha llegado al FTP desde el último envío y se borra el fichero del FTP

Tratamiento de la media





XML

MEDIA

VSN

MEDIA

DATOS

Etiqmedia

XML's

VSN

Transcodifica media
Interpreta XML recibido
Genera datos para enviar a Etiqmedia

Selecciona:
- Modelo de lenguaje
- Canal audio a analizar

Aplica AI basada en redes neuronales y otras tecnologías para procesado de la media
Genera XML con datos

Audio
- STT,
- diarización,
- Puntuación y capitalización
- Segmentación hablantes
- Reconocimiento entidades…
- Palabras clave
- Clasificación

Une información de Etiqmedia a los ficheros de media.
Quedan listos para control calidad

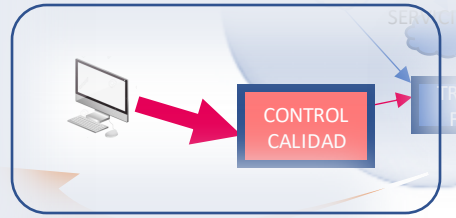
24 horas

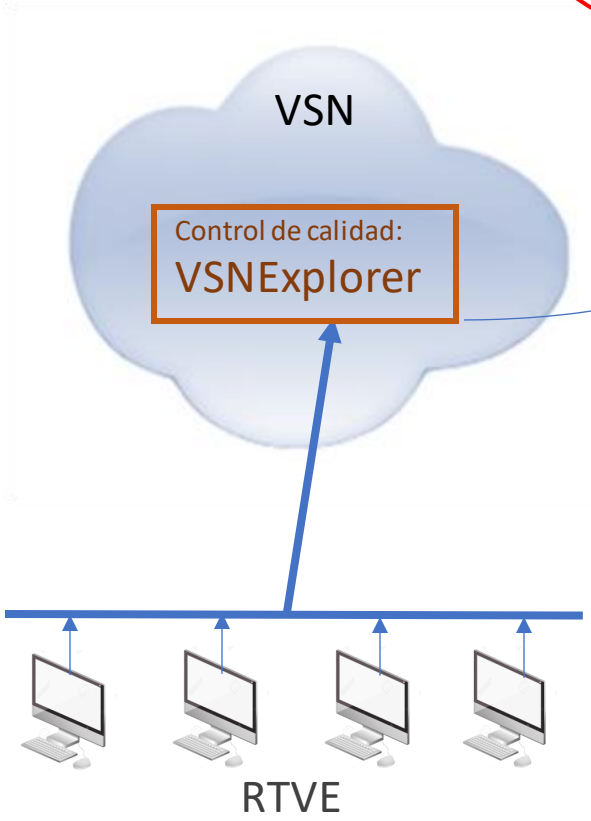
40 horas/dia (lunes a viernes)

Control de calidad

Herramienta fundamental

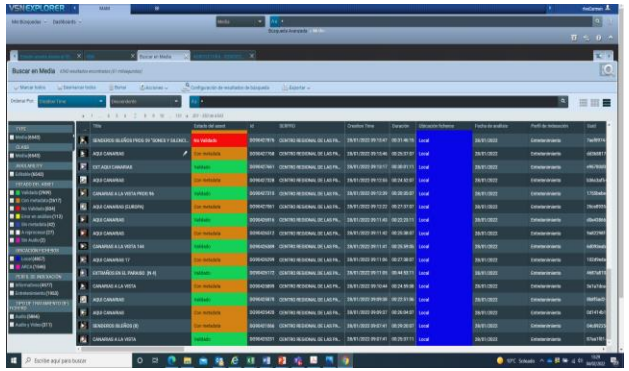
Conocer puntos fuertes y débiles
"Tunear" resultados a nuestra forma de trabajo
Permite garantizar la calidad de los metadatos que llegan a RTVE





Revisar y editar

Texto
Entidades,
palabras clave…
Validar o no resultados



Dashboard

Estado de los archivos: validados o no, con o sin metadata, con error…
Cálculos de WER, entidades, palabras clave, etc modificados…
Generar Excel
Hacer filtrados por fecha…

Envío metadatos a RTVE

VSN

Una vez al día en horario programado
Ficheros XML
Algoritmo automatiza envío
También todo lo generado por motor AI

RTVE
(FTP)

Sirve de enlace entre RTVE y VSN
Lo que llegue cada día se envía a ARCA

RTVE
ARCA

Incorpora en una pestaña dedicada a IA todos los datos recibidos
Almacena ficheros recibidos

Proceso	Automatizado	Automatizado

Material seleccionado para el proyecto

Contenidos con metadatos identificativos mínimos (título, serie, fecha de producción)

- Altas manuales realizadas por documentalistas
- Cargas automáticas con información procedente de los soportes.

Cobertura temática variada

- Deportes, cine, política, actualidad social, música, historia, literatura.

Cobertura cronológica muy amplia

- Desde los años 60 a 2010 (impacto sobre los procesos)

Distintos formatos de programas

- concursos, reportajes, magazines

Limitaciones en la selección

Limitaciones en la selección:

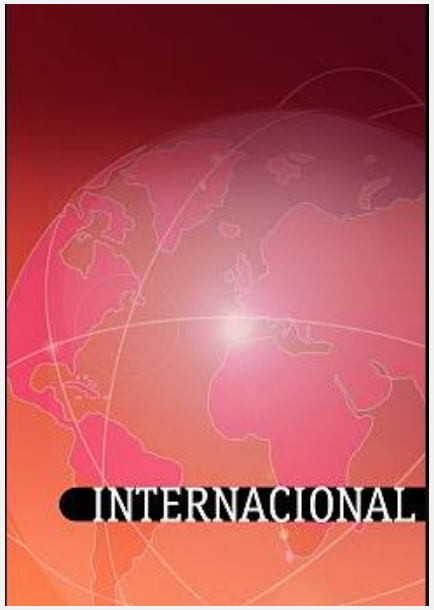
- Selección apriorística: se buscan contenidos con un alto % de audio.
- Información insuficiente o inexacta.
- Conocimiento de la tecnología adquirido en laboratorio (RTVE Iberspeech Challenge).

Nuevos criterios para la selección basados en

- Conocimiento adquirido en el proceso de validación
- La revisión previa de los contenidos antes de ser enviados.

Modelos de análisis definidos

- Proceso automático basado en los metadatos incluidos en el XML que se envía a VSN con el mp4.
- Tipo de contenidos:
  - Informativos
  - Programas
- Número de hablantes esperado
  - 1/10
  - 1/20
  - 1/50



Revista de actualidad + Cultura = Programas 1/50

Informativo = Informativos 1/50

Entrevista + Política = Informativos 1/10

Revista de actualidad + Medio ambiente = Programas 1/50

Modelos de validación

Automático

- Se aplica a programas con buenos resultados tras un testeo
- Programas bien estructurados, con predominio de locución profesional, en emisión actualmente o emitidos en los últimos 5 años.
- Nivel de confianza del 75%

Semi-automático

- Se revisa un 50% de los contenidos de una misma serie. Cuando se valida el 70% todo el "paquete" de horas queda validado.
- Algunos contenidos se automatizan tras un tiempo.
- Nivel de confianza en ARCA del 80%

Manual

- Se aplica a contenidos especialmente relevantes o complejos.
- Nivel de confianza del 90%

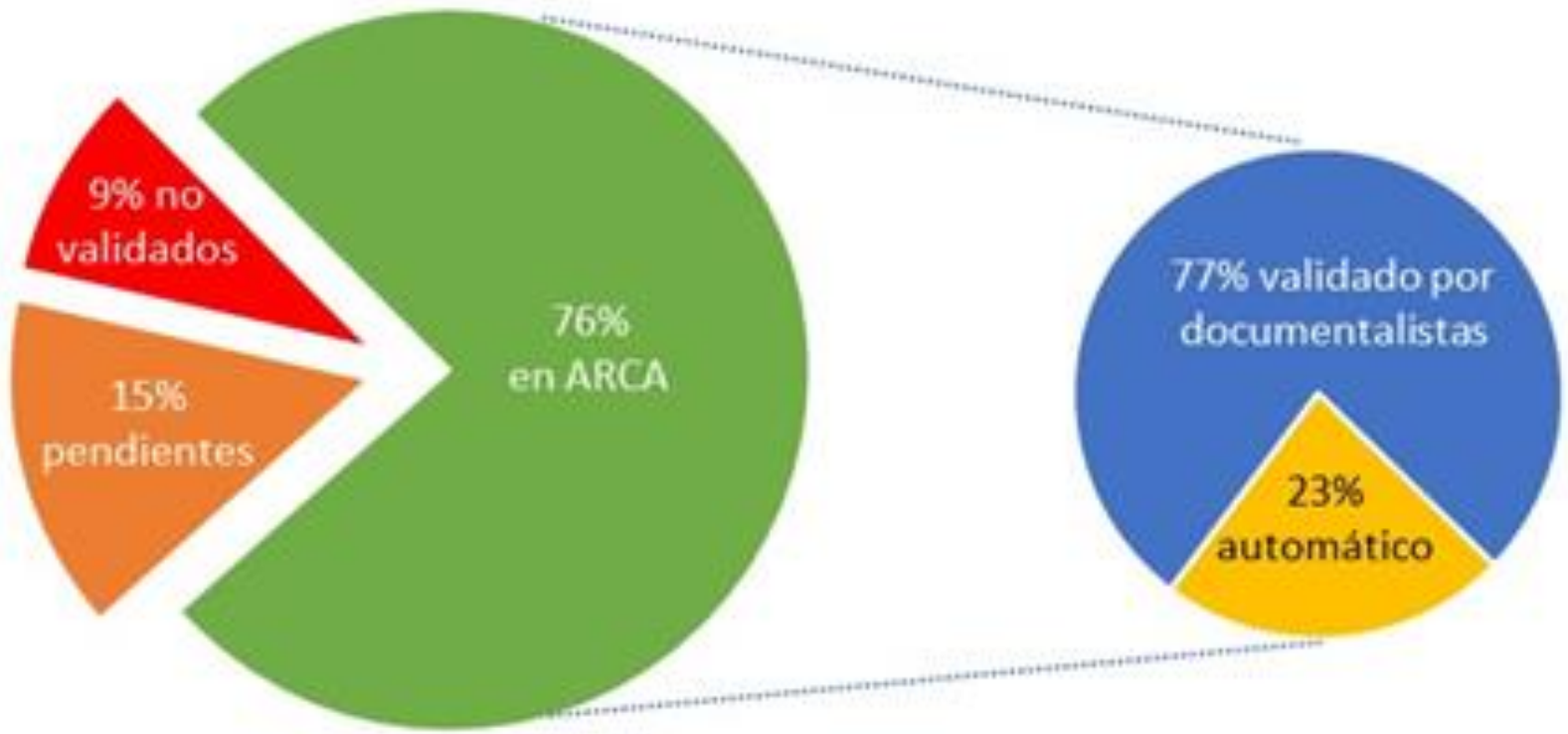
Criterios de validación

¿Es posible recuperar el contenido por sus elementos más relevantes?

En una búsqueda sobre un tema determinado ¿Sería pertinente la recuperación de este documento?

El tiempo que hay que invertir en corregir/añadir entidades, palabras clave, categorías no es excesivo.

Contenidos procesados: 13.224



Algunas conclusiones

La calidad de la transcripción está condicionada por factores como: ruido ambiente, música, claridad en la dicción, distintos acentos, idiomas.

La segmentación debe establecer fronteras claras entre hablantes y asignar de forma correcta cada segmento de voz a un hablante concreto. No es un problema resuelto.

No tenemos capacidad para influir en el rendimiento del sistema, solo en la selección previa del material que procesamos.

Se espera un mejor rendimiento del sistema en palabras clave, entidades y clasificación.

El estado actual de la tecnología marca los límites del proyecto.

Conclusiones

- Limitaciones relacionadas con la licitación https://licitaciones.rtve.es/
  - Contrato con el proveedor en el que se establecen las condiciones del servicio.
  - Funcionalidades e idiomas.
- Limitaciones relacionadas con la integración de los datos
  - Desarrollos en ARCA 3.0
- Limitaciones relacionadas con la tecnología

El futuro

- Selección parcial S2T.
- Integración del S2T con el minutado, subtitulos y reconocimiento de hablantes: recuperación multimodal
- Segmentación de subtítulos.
- Segmentación por contenido.
- Relación del S2T de cámaras masterizadas con las emisiones.
- Análisis de más de una pista de audio.
- Transcripción automática de contenidos en múltiples idiomas.
- Mayor detección de entidades de productos, fechas y eventos.
- Posibilidad de limitar entidades por documento, no por segmento.
- Mejora en la relevancia de entidades y palabras clave.
- Elaboración de resúmenes.
- Nuevos usos de los contenidos vinculados al fact-checking político y al periodismo de datos.