BACKEND PROCESSED MEDIA DATA
LIST OF MEDIA TO BE PROCESSED
LIST OF PROCESSED DATA FOR PROJECT

**IMPORTANT**: To generate current media results reports from the frontend like these.
SUBSCRIBER MEDIA RESULTS REPORT and
GENERAL PUBLIC MEDIA RESULTS REPORTS

Develop on our local computers **Digi62** to be able to administer the entire system.

**IMPORTANT**: Do not put any of the processed data or the backend software up on AWS Amazon Lightsail. Then put up CODENTV1C and CONDENTV2A. Connect the backend processed to http://44.231.94.92/ on AWS Amazon Lightsail. To generate Media Results Reports. **Please use C Code not C++**

**ARCHIVED PROCESSED MEDIA DATA**
**https://docs.google.com/document/d/1a6ZjjRkFuwxZwa8iwX3WcRPXg_Sq5MWzYr7cZulKJxA/edit?usp=sharing**

**TO PRESENT YOUR PROJECT:**

**BACKEND teams - needs to connect with teams working on frontend to generate media results reports on AWS Amazon Lightsail http://44.231.94.92/**

**MAKE A PLAN**: Divide the project up with other team members. **MERGE** your code a lot to digiclips henry's github master repository branch with comment notes linked to readme and version numbers.

Please contact Frontend Teams to present the project you have worked on.

**IMPORTANT NOTE**: Once these are fixed as we test it out we will let you know about more problems that need to be fixed. Put notes, comments into code and backup a lot to DigiClips Henry's github master branch. Auto Backup through ASUS Router Hard Drives on the local system. **30 days Auto delete** processed data from database and archived drives. Do a **production build** then put up on our local system for further testing and fixing problems. For your final presentation leave on the local system - chrome remote for your demonstration. **Allow 30 to 45 days for fixing problems**.

**IMPORTANT**: On Ubuntu never ever use # ROOT always use USER

**TEAMS WORKED ON PROJECT**: LEHIGH - FULLERTON

**Work on problems with buffer overflow and stop crashing error messages.**
**Make sure all processed media data is connected to the frontend media search options and email alerts frontend on digi-frontend computer and on AWS Amazon Lightsail to generate media results reports.**

NOTE: Language Translation is needed.

**VERY IMPORTANT**: Use https://www.ffmpeg.org/ to process the Television Web Streaming Radio Podcast Siriusxm data.

This is the what the backend processed media data should be to create the Media Results Reports from click here to explain what is needed> Options and Email Alerts frontend something like this: Rough Draft Sample Media Results Reports:

Click on the segments and go to the streaming Details Page - which is just a description of what is needed with the DETAILS STREAMING BUTTONS.
**Television**: http://www.digiclipsinc.com/digiview/TV_Image.png
**http://www.digiclipsinc.com/digiview/Sample%20of%20Stream_Television.htm**
**Radio**: http://www.digiclipsinc.com/digiview/Radio_Image.png
http://www.digiclipsinc.com/digiview/Sample%20of%20Stream_Radio.htm

NOTE: Allow 30 to 45 days to make it presentable to test and fix problems as we test it out. We will let you know about more problems that need to be fixed. Put notes, comments into code and backup a lot to DigiClips Henry's github branch. Put notes comments in your code put up on DigiClips github branch. Do a **production build** then put up on our local system for further testing and fixing problems. Then put the front end up on AWS Amazon Lightsail for your final presentation.

**BROADCAST**

**FREE TO AIR ANTENNA TELEVISION**

**WinTV-quad4 board used**
Free To Air Antenna https://www.hauppauge.co.uk/site/products/data_quadhd.html
For Hauppauge Support: **https://hauppauge.com/pages/support/support_linux.html**
Use software with the OS Operating System Ubuntu 22.04
**https://hauppauge.com/pages/support/support_linux.html**

NOTE: TV Stations should have network affiliations like ABC - CBS - NBC - FOX etc.

**TELEVISION CLOSED CAPTION TEXT**
**Close Caption text files should look like this but with all the different TV stations call letters. Import into mysql and into the archive directory. For now after 30 days auto delete from mysql docker container with snap and from archive drive. Should be variable and changed by Administration. The LeHigh2023 team has been working on this.**

KCNC -TV_2023_04_18_14-45-00_FTAAT-CC.ass
KCNC -TV_2023_04_18_14-45-00_FTAAT-CC-.vtt

## TELEVISION SPEECH TO TEXT

NOTE: In Stereo Left and right channel no video

**Speech to text in stereo for television should look like this but with all the different TV stations call letters. Import into mysql docker container with snap and into the archive directory. For now after 30 days auto delete from mysql and from archive drive. Should be variable and changed by Administration. Working on with the LeHigh Team 2023.**

KCNC -TV_2023_04_17-15-25-05-FTAAT-TST-Left.txt
KCNC -TV_2023_04_17-15-25-05-FTAAT-TST-Right.txt

## TELEVISION FREE TO AIR ANTENNA VIDEO TEXT RECOGNITION

**Video to text television should look like this but with all the different TV stations call letters. Import into mysql docker container with snap and into the archive directory. For now after 30 days auto delete from mysql and from archive drive. Should be variable and changed by Administration. Worked on by the Fullerton Team 2022.**

KCNC -TV_2023_04_18_15-00-03-VTR.txt

## IMAGE RECOGNITION

NOTE: This needs to be planned developed and worked on

## LIP READ TO TEXT

NOTE: This needs to be planned developed and worked on

## IMPORTANT MAKE PROJECT PRESENTABLE:

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

## WEB STREAMING TV

NOTE: Capture process the web TV streaming data.
Close caption text - Speech to Text stereo - Video Text Recognition - Image Recognition
Use for development on Digi62 computer:
WISH https://www.wishtv.com/live-stream/
CNBC http://www.freeintertv.com/view/id-2566
NOTE: Use this one to develop then through administration add on more later.

**Import into mysql and into the archive directory. Start with 3 days then auto delete from mysql and from archive drive. Should be variable and changed by Administration.**

WISH https://www.wishtv.com/live-stream/
**CNBC http://www.freeintertv.com/view/id-2566**
**Open Source Software** that might help: https://obsproject.com/download#linux

**CLOSE CAPTION**
CNBC_2023_04_18_14-45-00_CC.ass
CNBC_2023_04_18_14-45-00_CC.vtt

**Import into mysql docker container with snap and into the archive directory. For now after 30 days auto delete from mysql and from archive drive. Should be variable and changed by Administration.**

**IMPORTANT MAKE PROJECT PRESENTABLE**:

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

**SPEECH TO TEXT**
CNBC_2023_04_17-15-25-05_WST-SST-Left.txt
CNBC_2023_04_17-15-25-05_WST-SST-Right.txt

**Import into mysql docker container with snap and into the archive directory. For now after 30 days auto delete from mysql and from archive drive. Should be variable and changed by Administration.**

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

**VIDEO TEXT RECOGNITION**
CNBC_2023_04_18_15-00-03--VTR.txt

**Import into mysql docker container with snap and into the archive directory. For now after 30 days auto delete from mysql and from archive drive. Should be variable and changed by Administration. Was worked on by the Fullerton 2022-23 Team**

**IMPORTANT MAKE PROJECT PRESENTABLE**:

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results**

**reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

**RADIO SPEECH TO TEXT**
Radio AM FM - Podcast - Sirius Radio

**Radio processed speech to text should look like this but with all the different Radio stations call letters. Then divided up into AM and FM. Import into mysql and into the archive directory. For now after 30 days auto delete from mysql docker container with snap and from archive drive. Connect the processed data through http://44.231.94.92/ on AWS Amazon Lightsail to generate Media Results Reports. Should be variable and changed by Administration. Worked on by the LeHigh Team**

KOA AM _Radio_2023_04_17-15-25-05-RST-Left.txt
KOA AM _Radio_2023_04_17-15-25-05-RST-Right.txt
KOA AM _Radio_2023_04_17-15-25-05-RST-Left.wav
KOA AM _Radio_2023_04_17-15-25-44-RST-Right.wav

KOA FM _Radio_2023_04_17-15-25-05-RST-Left.txt
KOA FM _Radio_2023_04_17-15-25-05-RST-Right.txt
KOA FM _Radio_2023_04_17-15-25-05-RST-Left.wav
KOA FM _Radio_2023_04_17-15-25-44-RST-Right.wav

**IMPORTANT MAKE PROJECT PRESENTABLE**:

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

**PODCAST RADIO SPEECH TO TEXT**
NOTE: Sample of the way I think the way the files should look like and with the Podcast Station. **Import into mysql docker container with snap and into the archive directory. For now after 30 days auto delete from mysql and from archive drive. Connect the processed data through http://44.231.94.92/ on AWS Amazon Lightsail to generate Media Results Reports. Should be variable and changed by Administration.**

Text Name

DAILY WIRE _Podcast_2023_04_17-15-25-05-PRST-Left.txt
DAILY WIRE _Podcast_2023_04_17-15-25-05-PRST-Right.txt
DAILY WIRE _Podcast_2023_04_17-15-25-05-PRST-Left.wav
DAILY WIRE _Podcast_2023_04_17-15-25-44-PRST-Right.wav

**IMPORTANT MAKE PROJECT PRESENTABLE**:

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

**SIRIUSXM RADIO SPEECH TO TEXT**

NOTE: Sample of the way I think the way the files should look like and with the Podcast Station. **Import into mysql and into the archive directory. For now after 30 days auto delete from mysql docker container with snap and from archive drive. Connect the processed data through http://44.231.94.92/ on AWS Amazon Lightsail to generate Media Results Reports. Should be variable and changed by Administration.**

**BLOOMBERG RADIO**
**https://www.siriusxm.com/player/channel-linear/bloomberg-radio/27adcf33-1681-9f00-2461-2129a2471abf**

Text Name

BLOOMBERG RADIO _SiriusxmRadio_2023_04_17-15-25-05-SRST-Left.txt
BLOOMBERG RADIO _SiriusxmRadio_2023_04_17-15-25-05-SRST-Right.txt
BLOOMBERG RADIO _SiriusxmRadio_2023_04_17-15-25-05-SRST-Left.wav
BLOOMBERG RADIO _SiriusxmRadio_2023_04_17-15-25-44-SRST-Right.wav

**IMPORTANT MAKE PROJECT PRESENTABLE**:

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

**INTERNET**
Webmedia: Newspapers - Magazine - Social Media - Blogs
Note: Scraping the web deleting cookies
https://www.octoparse.com/blog/10-best-open-source-web-scraper

**IMPORTANT MAKE PROJECT PRESENTABLE**

**Connect this processed media data from the frontend on both digi-frontend computer and on AWS Amazon Lightsail through http://44.231.94.92/ to generate media results reports from options and email alerts SUBSCRIBER MEDIA RESULTS REPORT and GENERAL PUBLIC MEDIA RESULTS REPORTS**

## WEB STREAMING TELEVISION

Use for developing

**Develop using** WISH TV https://www.wishtv.com/live-stream/ or
CNBC http://www.freeintertv.com/view/id-2566

ADD ON MORE STATIONS: https://www.digiclipsinc.com/usatvstreamainmedia.htm

Data processed close caption text - speech to text - video text recognition - image recognition

YOUTUBE  Use WBBM TV, a Chicago TV Station. Data processed close caption text - speech to text - video text recognition - image recognition to generate media results reports.

==========

APACHE NUTCH
SCRAPY
APIFY
PYSPIDER
STORMCRAWLER

Web TV Streaming test data sample of what they should look like

CNBC_Webstream_2023_04_17-15-25-05-ST-Left.txt
CNBC_Webstream_2023_04_17-15-25-05-ST-Right.txt
CNBC_Webstream_2023_04_17-15-25-05-VTR.txt
CNBC_Webstream_2023_04_17-15-25-05-CC.txt

YOUTUBE

**Work on problems with buffer overflow and stop crashing.**

**IMPORT THE DATA**

Into mysql relational database to create the media results reports

**BACK END**:

OPEN SOURCE USE THIS FOR DEVELOPMENT AND EXPERIMENT

**PROCESSING DATA TV AND RADIO**

www.ffmpeg.org
Fix processing data with buffer overload
And more
Relational Database: https://www.mysql.com/
Open Source Relational Databases: https://kinsta.com/blog/open-source-database/
Snap: https://snapcraft.io/docs/docker-to-snap
Xmedia: https://www.xmedia-recode.de/en/

**Speech to text**
https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-container-howto?tabs=stt%2Ccsharp%2Csimple-format
https://fosspost.org/open-source-speech-recognition/

VOSK
PaddleSpeech
OTTER
WHISPER
ATHENA

**Video text recognition**
https://azure.microsoft.com/en-us/blog/text-recognition-for-video-in-microsoft-video-indexer/
https://www.hitechnectar.com/blogs/open-source-ocr-tools/#:~:text=A9T9-,Tesseract,available%20open%2Dsource%20systems%20available.

TESSERACT
A9+9

**ARCHIVED DATA PROCESSED**

Television free to air cable streaming TV Media closed caption text - speech to text - video text recognition - image recognition
Radio podcast sirius radio speech to text
Newspapers - magazines - social media blogs scrap the web
Compress data in hi quality save on archived drives

**Image Recognition**
**Needs worked on**
https://learn.microsoft.com/en-us/azure/azure-video-indexer/video-indexer-overview
https://opensource.com/article/18/5/getting-started-luminoth
https://cmusatyalab.github.io/openface/
https://nordicapis.com/7-best-image-recognition-apis/
https://www.g2.com/categories/image-recognition/free

CLOUDVISION API

AMAZON REKOGNITION
CLARIFAI
CVEDIA
ROBOFLOW

**Lip Read To Text**
**https://learn.microsoft.com/en-us/answers/questions/1183790/code-sample-for-lip-sync**

**Auto Delete Cookies**

**CONNECTING THE FRONTEND TO THE PROCESSED DATA ON OUR LOCAL SYSTEM**
https://docs.aws.amazon.com/lightsail/latest/userguide/amazon-lightsail-connecting-to-your-mysql-database.html

**Administration**

**For now start with 3 days with websteaming television podcast sirius radio with just 3 Denver Radio stations and TV free to air antenna 30 days delete everyday from mysql and from the archive drive.  Connect the processed data through http://44.231.94.92/ on AWS Amazon Lightsail to generate Media Results Reports. Should be variable and changed by Administration. Administration should be able to add on more media.**