

# GUIDELINES FOR BROADCAST AUDIO COLLECTION

Version 3.0

August 13, 2010

Linguistic Data Consortium

<http://www ldc.upenn.edu/Projects/GALE>

## Introduction

The goal of the GALE Broadcast Collection task is to collect Arabic, Chinese and English broadcast news (BN) and broadcast conversation (BC) programming. BN programming consists of “talking head” style broadcasts, i.e., generally one person reading a news script. BC programming is more interactive and includes talk shows, interviews, call-in programs and roundtables. A program’s classification as BN or BC is intended to reflect that program’s dominant genre. Both genres may occur within a single program.

The Linguistic Data Consortium (LDC) maintains a local broadcast collection at its Philadelphia facilities and manages remote collection sites for the GALE broadcast audio collection. The remote collection sites are Hong Kong University of Science and Technology (HKUST), Hong Kong, Republic of China (Chinese); Medianet, Tunis, Tunisia (Arabic); and MTC, Rabat, Morocco (Arabic). The combined local and outsourced broadcast collection supports GALE at a rate of approximately 300 hours per week of programming from 40 broadcast sources.

LDC has designed a portable broadcast collection platform for remote broadcast collection that is used by HKUST and Medianet for a portion of their outsourced collections. The portable broadcast collection platform is described in further detail below.

## LDC’s Local Broadcast Collection System

A full description of LDC’s broadcast collection system that includes the collections database, hardware schema, utilities and data management functions can be found in LDC Broadcast Collection System Documentation – Version 1.0, [http://projects ldc.upenn.edu/gale/task\\_specifications/LDC Broadcast System Documentation v1.0.pdf](http://projects ldc.upenn.edu/gale/task_specifications/LDC_Broadcast_System_Documentation_v1.0.pdf) Below is an overview of the system, the sources recorded and the configuration of the recording laboratory.

LDC collects broadcast data from a variety of programs in the target languages emanating from the following sources: 2M TV, Abu Dhabi TV, Al Alam News Channel, Al Arabiyah, Al Iraqiyah, Aljazeera, Al Ordiniyah, Dubai TV, Kuwait TV, Lebanese Broadcasting Corp., Nile TV, Oman TV, Saudi TV, SCOLA Foreign Language Network (SCOLA), Syria TV, Tunis TV (Arabic); Beijing TV, China Central TV (CCTV), Dragon TV, Fujian TV, Guangdong TV, Jiangsu TV, New Tang Dynasty TV (NTDTV) and Phoenix TV, Zhejiang TV (Chinese); and CNN/CNN Headline News and MSNBC/NBC

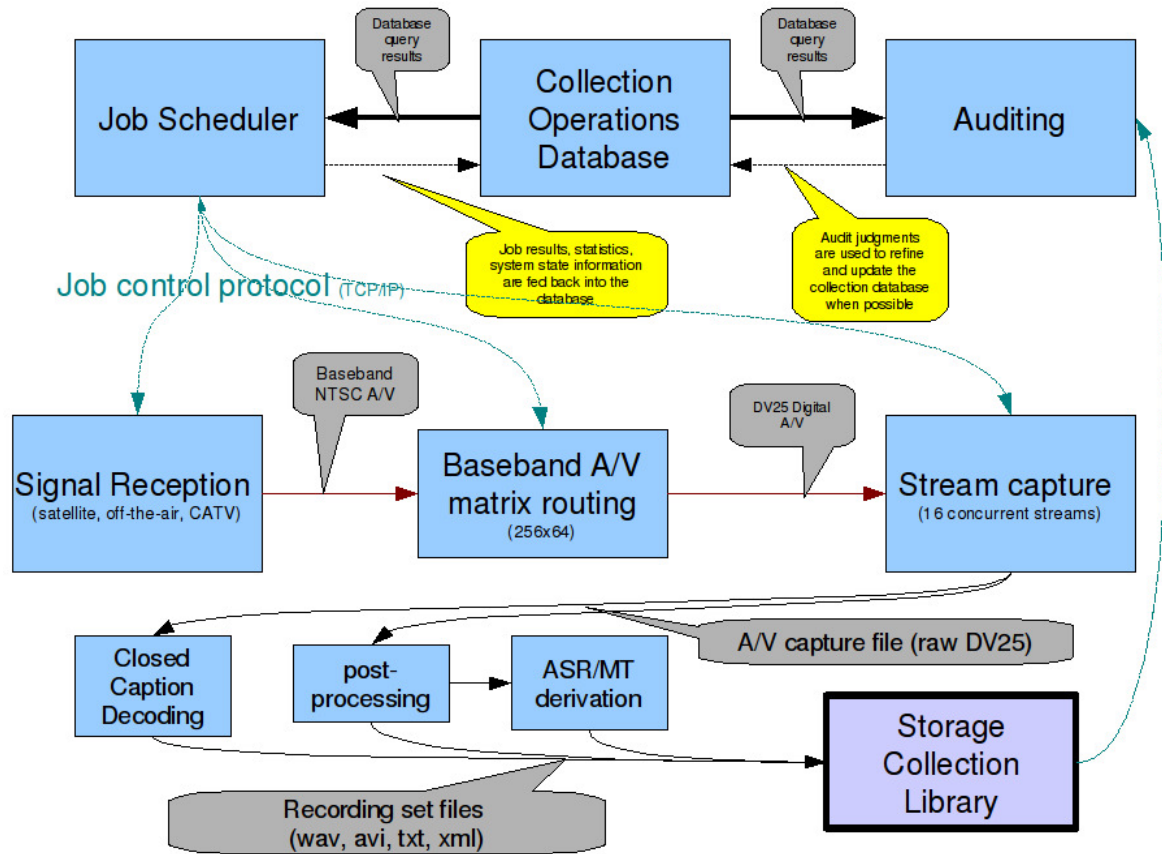
(English). Sources and programs may change from time to time depending on availability, project requirements or other factors.

LDC maintains six satellite dishes that provide access to C-Band, Ku-Band, DirecTV and DISH Network programming. A 3.7 meter solid dish and a 3.1 meter sectional dish are installed on movable, horizon-to-horizon mounts and are connected to computer-controlled dish movers. Four small fixed direct broadcast satellite dishes (DBS) provide access to DirecTV and Dish Network. LDC currently operates one Wegener MPEG-1 receiver for SCOLA multilingual broadcasts, two dedicated DVB-S receivers for CCTV and Phoenix TV programming, one Chapperal M100+ receiver, and six Pansat DVB/MPEG-2 receivers for free to air broadcasts and wild feeds on both C and KU bands. There are eight Dish Network receivers which provide the capability to receive all of Dish Network's domestic and international programming and one DirecTV receiver for domestic U.S. programming. A Dressler active shortwave antenna together with an AOR wideband antenna cover the entire range of the electromagnetic spectrum used for speech communication. The University's Penn Video Network is an additional broadcast source and provides local, national and international television programming.

A control computer coordinates the activities of all satellite dishes and receivers and CATV tuners/demodulators routing signals via two Knox AV matrix switches (64 inputs / 32 outputs) and sixteen distribution amplifiers to eight Linux-based recording nodes. Each recording node is capable of simultaneously capturing two streams of DV25 digital video+audio direct to local disk. The broadcast collection system also includes substantial, flexible monitoring capabilities via an integrated LCD monitoring matrix (nine separate video monitors, 4 channels of audio). A gigabit Ethernet switch connects all broadcast recording hardware to LDC's static storage and backup facilities.

As a program is recorded, the analog audio and video pass through the A/V matrix switch to a Canopus ACEDVio analog to DV converter. The digitized DV25 stream contains DV video (intraframe compression, 4:1:1 color space, 720x480 frame resolution, 30fps) and linear PCM audio (stereo, 48kHz, 16 bit/sample). The raw DV25 stream is captured to disk using dvgrab (<http://www.kinodv.org>). Once a program has been recorded, the linear PCM audio tracks are extracted from the raw stream. The extracted audio (48kHz, 16bit, stereo) is downsampled to 16kHz and split into separate tracks. Then, the raw stream is converted to an MPEG-4 avi (30fps, 720x480) with a target bitrate of 1Mbps (896Kbps video, 128Kbps audio). The extracted audio and the MPEG-4 avi are uploaded to LDC's fileserver. The audio files are saved as .wav files. This process is depicted graphically in the LDC Broadcast Collection System Block Diagram below.

### LDC Broadcast Collection System Block Diagram



Following is a list of the broadcast sources LDC receives by the delivery modes set forth above:

**CNN, CNN Headline News, MSNBC (alternate source), NBC --** Local Philadelphia CATV sources delivered to LDC as NTSC modulated sources over analog CATV

**Abu Dhabi TV, Al Arabiyah, Aljazeera, LBC, MSNBC, Nile TV (SCOLA), Oman TV –** Dish Network sources delivered to LDC as MPEG-2 TS over DVB-S with conditional access

**2M TV, Al Alam News Channel, Al Iraqiyah, Dubai TV, Kuwait TV, Nile TV (alternate source), NTDTV, Oman TV (alternate source), Saudi TV, Syria TV –** Free-to-air sources delivered to LDC as MPEG-2TS over DVB-S in the clear on Galaxy 25 (97.0°W)

**CCTV (Galaxy IIC, 95.0°W), Phoenix TV (Galaxy IIC, 95.0°W), SCOLA (alternate source for Dubai TV, Wegener MPEG-1, qpak, Galaxy 25 (97.0°W)) –** Other conditional access satellite DVB-S sources

A photograph of LDC's broadcast collection system is shown below.



Shown from left to right are:

1. **Left rack (top to bottom):** Coship receivers for Arabic language programming (5); SCOLA receiver (1); Dish network receivers (8); Pansat/Coolsat receivers for free-to-air broadcasts (4); CCTV and Phoenix TV receivers (2); Penn Video Network receivers (4); distribution amplifiers and associated supplies.
2. **Middle rack (top to bottom):** Coolsat receivers for Arabic language programming (2); DirecTV receiver (1); monitor bank showing programs currently or last recorded; larger monitor at top center can be manually tuned to test/check programming; AV Matrix (1), routes audio/video output from receivers to recording nodes and monitors; Control devicemasters (2) serve as communication layer between the AV matrix, closed caption decoder and main broadcast server (latter not shown); Hubcap closed caption decoders (17).

3. **Right rack (top to bottom):** Recording nodes (2) used for testing and periodic channel surveys; recording nodes (6) for current recordings; switch box (1) used to connect to a local network; universal power supplies (3) for back-up power.

### **ASR Output**

LDC has integrated three ASR client systems into its daily collection process for the duration of the GALE program. They were developed by GALE research sites BBN Technologies (BBN), International Business Machines (IBM) and SRI International (SRI). ASR output is automatically generated on the BBN and IBM systems for all locally-collected Arabic and Chinese audio data. The text output is used for downstream data selection. Audio data from LDC's remote broadcast sites is not processed daily, but is run as needed as part of the data selection process. Data is processed on the SRI system by request.

The IBM and SRI applications are installed on local machines at LDC and administered by LDC staff. Data is processed on the BBN system via a connection to remote servers which necessitates an additional file anonymization step.

Locally collected broadcasts are automatically pooled onto a centralized server as they are recorded and processed. This server then supplies the extracted audio portions to the different ASR systems which run daily cronjobs to generate ASR out put for the previous day's broadcasts.

Since each system has different requirements, wrapper scripts were created by LDC to preprocess the input .wav audio file, to run the actual ASR application on each file and to perform additional postprocessing and formatting before copying the output back to the collections server.

### **Closed Caption (CCAP) Output**

Closed caption (CCAP) output for all English-language programs is generated at the time of transmission by the "SoftTouch Hubcap" system which decodes North American line-21 closed captions, the only CCAP system used in the United States. If CCAP output cannot be collected at the time of transmission because of technical problems in the transmission or because of problems with LDC's broadcast collection system, CCAP cannot be "re-generated" after transmission.

### **Web Broadcast Collection**

LDC also collects broadcast audio and transcripts from selected websites that are included in GALE broadcast releases after appropriate review and auditing. Websites from which such data are collected include Aljazeera (Arabic) and New Tang Dynasty TV (Chinese).

## **Remote Broadcast Collection**

HKUST collects Chinese BN and BC programming using its internal recording system and a portable broadcast collection platform designed by the LDC and installed at HKUST in GALE Phase 1. Among the sources collected by HKUST for GALE are Anhui TV, Beijing TV, China Central TV, Dongfang TV, Fujian TV, Hubei TV, Jiangsu TV and Voice of America (VOA) Mandarin.

Medianet collects Arabic BN and BC programming that contains a range of Arabic dialects from across the Gulf region using its internal recording system and LDC's portable broadcast collection platform installed in 2008. Among the sources collected by Medianet are Abu Dhabi TV, Al Arabiyah, Al Baghdadya, Al Fayhaa, Al Forat, Al Hiwar, Al Iraqiyah, Al Manar, Al Ordiniyah, Al Sharqiya, Bahrain TV, Dubai TV, Kuwait TV, Oman TV, Qatar TV, Palestine Satellite Channel, Saudi TV and Tunis TV.

MTC collects Arabic BN and BC programming from Al Baghdadya, Alhurra, Al Maghribia, Arabiia, Radio Sawa and Yemen TV using its internal collection system.

## **Portable Broadcast Collection Platform**

LDC's portable broadcast collection platform is a TiVO style DVR system capable of recording two streams of AV material simultaneously; it supports analog CATV (NTSC and PAL) and FTA DVB-S satellite programming, and is capable of operating outside of the United States. It has a small footprint and can be transported as carry-on luggage.

The portable broadcast collection platforms delivered to HKUST and Medianet weigh less than 30 pounds, have a footprint no larger than 60cm x 60cm x 10cm and contain scheduling software, diagnostic tools and remote control functionality. The components of the portable broadcast collection platform delivered to Medianet are shown in the photograph below.



The portable platform and the LDC collection system share the same code base and rely on a modular, unified hardware specification. Improvements in the main collection platform therefore translate into benefits for the portable platform.

### **Quality Control**

With respect to LDC's local collection, the Manager of the Recording Lab and the Recording Lab Programmer monitor the recording equipment and storage capacity to ensure the quality and accessibility of the broadcast collection. In addition, the Broadcast Collection Manager and the Recording Lab Manager monitor broadcast schedules for the target programming to keep the number of unusable recordings to a minimum. A 25% error rate is assumed, and additional programming (approximately 25% over the target amount) is collected as an offset.

Remote collection sites make any infrastructure changes to receive the target programming in consultation with LDC. LDC works with remote sites to determine recording and delivery file formats, to select suitable programming and to establish and maintain broadcast schedules. Remote collection sites send periodic collection reports to LDC which LDC monitors to ensure that the remote collections are proceeding as planned.

Remote collection sites deliver recording files to LDC by ftp server according to a schedule determined in consultation with LDC. LDC downloads and processes the outsourced files and makes them available for release to GALE sites and for downstream data selection.

## **Distribution**

Broadcast audio data is distributed on a hard disk drive to GALE registered users in accordance with the GALE release schedule.