

Ingest Unstructured Data / Output Structured Evidence In laymen's terms DVD/LTO/VHS to Cold Storage

Preface: This document starts with a more technical conversation as it relates to the distinction between unstructured data and structured data. Then it ends with real world applications without the technical "propeller head" jargon.

(Technical Part)

What is the difference between structured and unstructured data?

Structured data is highly organized and formatted so that it is easily searchable in relational databases. Structured data can be thought of as records (or transactions) in a database environment; such as rows in a table of a SQL database. Unstructured data has no predefined format or organization, making it much more difficult to collect, process, and analyze. Think of unstructured data like a large junk drawer of all your data, or data that does not live in a relational database management system (RDBMS). Unstructured data is unmanageable as evidence.

There is no "rules logic" as to whether data is structured or unstructured. Unstructured data just happens to be in greater abundance than structured data.

Examples of unstructured data are:

- Rich media, media and entertainment data, surveillance data, geo-spatial data, audio, weather data
- Document collections, invoices, records, emails, productivity applications

- Internet of Things (IoT), sensor data, ticker data
- Analytics, machine learning, artificial intelligence (AI)

There are notable differences between structured and unstructured data to be aware of when dealing with any of the data types. The following table will help compare the two types of data based on factors such as data sources, data storage, internal structure, data format, scalability, usage, and more.

Structured data	Unstructured data
Structured data is quantitative data that consists of numbers and values.	Unstructured data is qualitative data that consists of audio, video, sensors, descriptions, and more.
Structured data is used in machine learning and drives machine learning algorithms.	Unstructured data is used in natural language processing and text mining.
Structured data is stored in tabular formats like excel sheets or SQL databases.	Stored as audio files, videos files, or NoSQL databases
Structured data has a pre-defined data model.	Unstructured data does not have a pre-defined data model.
Structured data is sourced from online forms, GPS sensors, network logs, web server logs, OLTP systems, and the like.	Unstructured data is sourced from email messages, word-processing documents, pdf files, and so on.
Structured data is stored in data warehouses	Unstructured data is stored in data lakes
Structured data requires less storage space and is highly scalable.	Unstructured data requires more storage space and is difficult to scale.

Semi-structured data

Semi-structured data is what is sounds like...data that lies midway between structured and unstructured data. It does not have a specific relational or tabular data model but includes tags (metadata) and semantic markers that scale data into records and fields in a dataset.

Common examples of semi-structured data are JSON and XML. Semi-structured data is more complex than structured data but less complex than unstructured data. It is also relatively easier to store than unstructured data and bridges the gap between the two data types.

Metadata - the master data

Metadata is often used in big data analysis and is a master dataset that describes other data types. It has preset fields that contain additional information about a specific dataset. Metadata has a defined structure identified by a metadata markup schema that includes metadata models and metadata standards. It contains valuable details to help users better analyze and manage data items and make informed decisions.

For example, an online article can display metadata such as a headline, a snippet, a featured image, image alttext, slug, and other related information. This information helps differentiate one piece of content from other similar pieces of content on the web. Metadata is, therefore, a handy descriptive method in which easy searches are executed.

(Non "propeller head" jargon)

Real-World Problems

Law Enforcement Agencies are faced with challenges regarding their archived evidence. Terabytes and even Petabytes of unstructured data exist, resulting in exceptionally large "junk drawers." These junk drawers are comprised of servers/storage and secondary devices including DVDs, VHS tapes, thumb drives, LTO tapes, among others. With a better understanding of un/semi/structured data, the task of finding a specific case or file becomes akin to the proverbial "needle in a haystack."

Another risk faced by the LEAs, Prosecutors, and Courts is keeping archived files on secondary devices and stored in warehouses where there is a real risk of natural disaster and/or deterioration of the media. The risk of flooding, fires, and earthquakes are an all too real issue as it has already caused evidence to be destroyed for many LEAs. The deterioration of media is also hitting agencies today since most of the secondary devices used are from years ago and even decades ago. Take a look at the lifespan image in the post linked here (LinkedIn Post).

Another issue is the technology to playback the secondary media is becoming obsolete. When was the last time you purchased a computer/laptop and found a CD/DVD player built in? VCRs are almost impossible to find and are now expensive collectors' items. Even USB ports have changed requiring different adaptors to support thumb drives or external Hard Disk Drives. LTO tape manufacturers have come and gone... who remembers ZIP drives?

So, the ultimate question becomes, "how do we take the unstructured data and turn it into structured and how do we archive our secondary devices to cold storage?"

Real-World Solutions

Nearly every organization has archiving concerns. Many have a budget to tackle the effort, but most do not have people and resources available to process content. There are options like hiring additional staff, or bringing in temporary help, but that is not always possible and potentially a lot to manage. Another option is to outsource the services required to a certified third-party. A certified person, with a background check, can come onsite, provide the secondary device hardware (i.e., VCR, LTO, DVD publisher by Rimage) needed to take digital evidence and run it through a secure, hash verified, and encrypted process. This process can optionally include proprietary video conversion, transcription services, and can send the files directly to your desired storage and or your backend software (i.e., DEMS, CMS, RMS).

Output structured Files

DWS's Data-Central becomes the bridge for the transformation of data. **DC** takes all of the data from DVD, VHS tapes, etc. and adds metadata, proprietary video conversion, user notes/priority settings, and adds it to a database. Once ingested into the database, all of the metadata, file names, and folder structures can be searched with a simple wildcard search. Complex nested filters are created using a simple UI to sort through hundreds of thousands of files in seconds. The proprietary videos will automatically be converted to standard .mp4 files (working products). These files are automatically to 3rd party enrichment/analytic tools for an even deeper source of meaningful metadata. Audio/video files can be transcribed allowing for a full contextual search of the file.

Once all of the files have been ingested into **Data-Central**, an AES 256 encrypted **Case-Pak** is generated. The database, along with the files, metadata, and working products in the encrypted **Case-Pak** are output to active or cold storage.



The **Case-Pak** can be output back onto a DVD, HDD, or thumb drive if desired.



DWS's Data-Central addresses proper and affordable solutions for evidence handling. **DWS's Data-Central** is middle-ware specifically written with digital evidence processing and security complications in mind. The **DWS** team not only has the software, but the experience to deal with these issues properly and efficiently. If you are challenged by current technology, budgets, and time-consuming manual processes and would like a demo or more information on **DWS's Data-Central** product, please <u>CLICK HERE</u>.

Dynamicworkflowsolutions.com