Name: John J. Pretz, **Course:** AI574, **Professor:** Prof. Youakim Badr, **Semester:** Fall 2025

Grateful Dead Concert Metadata Scraper (1990 - Sampled 1500 Items)

This script retrieves metadata about Grateful Dead concerts from the year 1990 using the Internet Archive Scraping API. To keep the dataset manageable, it saves only the first 1500 results, even though the archive contains more items for that year. At the end, the script prints a short summary (earliest date, latest date, most downloaded show).

Data Source:

- Internet Archive, "Grateful Dead Collection" URL: https://archive.org/details/GratefulDead

API Documentation:

- Internet Archive Search & Scraping API https://archive.org/help/aboutsearch.htm

Example Items (used for metadata reference):

- Grateful Dead Live at Capital Centre on 1990-03-14 https://archive.org/details/gd1990-03-14.Nak300CP4.Fitzy.Keo.125852.Flac1644
- Grateful Dead Live at Deer Creek Music Center on 1990-07-18 https://archive.org/details/gd1990-07-18.schoeps.miller.117361.flac16

Credits:

- Internet Archive for providing open and accessible concert recordings and metadata through their API.
- Code structure adapted from Archive.org API usage examples.

In [1]:
```
import requests
import json
import time
from datetime import datetime

def fetch_metadata(limit=1500):
    endpoint = "https://archive.org/services/search/v1/scrape"
    q = 'collection:GratefulDead AND year:1990'
    fields = [
        "identifier", "title", "description", "creator", "year",
        "addeddate", "collection", "mediatype", "item_size",
        "subjects", "publicdate", "downloads", "files"
    ]
    params = {"q": q, "fields": ",".join(fields), "count": 100}
```

```
    cursor = None
    all_items = []

    while len(all_items) < limit:
        if cursor:
            params["cursor"] = cursor

        response = requests.get(endpoint, params=params)
        response.raise_for_status()
        data = response.json()

        items = data.get("items", [])
        all_items.extend(items)
        cursor = data.get("cursor")

        print(f"Fetched {len(all_items)} items so far...")

        if not cursor:
            break
        time.sleep(1)

    return all_items[:limit]

def save_to_file(items, filename="grateful_dead_1990_sample.json"):
    with open(filename, "w", encoding="utf-8") as f:
        json.dump(items, f, indent=2)

def summarize(items):
    dates, most_downloaded, max_downloads = [], None, -1

    for item in items:
        pubdate = item.get("publicdate")
        if pubdate:
            try:
                dates.append(datetime.fromisoformat(pubdate.replace("Z", "")))
            except ValueError:
                pass
        downloads = item.get("downloads", 0)
        if downloads is not None and downloads > max_downloads:
            max_downloads, most_downloaded = downloads, item

    earliest = min(dates).strftime("%Y-%m-%d") if dates else "N/A"
    latest = max(dates).strftime("%Y-%m-%d") if dates else "N/A"

    print("\n📊 Dataset Summary (1990 Sample)")
    print("--------------------------------")
    print(f"Total items analyzed: {len(items)}")
    print(f"Earliest archive public date: {earliest}")
    print(f"Latest archive public date: {latest}")
    if most_downloaded:
        print(f"Most downloaded show: {most_downloaded.get('title', 'Unknown')}")
        print(f"Downloads: {max_downloads}")
        print(f"Identifier: {most_downloaded.get('identifier')}")

def main():
    items = fetch_metadata(limit=1500)
```

```
    print(f"\n✅ Total items fetched (sampled): {len(items)}")
    save_to_file(items)
    summarize(items)

if __name__ == "__main__":
    main()
```

```
Fetched 100 items so far...
Fetched 200 items so far...
Fetched 300 items so far...
Fetched 400 items so far...
Fetched 500 items so far...
Fetched 600 items so far...
Fetched 700 items so far...
Fetched 800 items so far...
Fetched 900 items so far...
Fetched 1000 items so far...
Fetched 1100 items so far...
Fetched 1200 items so far...
Fetched 1300 items so far...
Fetched 1400 items so far...
Fetched 1500 items so far...

✅ Total items fetched (sampled): 1500

📊 Dataset Summary (1990 Sample)
--------------------------------
Total items analyzed: 1500
Earliest archive public date: 2005-06-02
Latest archive public date: 2025-06-18
Most downloaded show: Grateful Dead Live at Capital Centre on 1990-03-15
Downloads: 86742
Identifier: gd1990-03-15.28293.sbeok.shnf
```

In [ ]: