Linked Open Data at the Anthology Film Archives

Level I Planning Grant Application Funding level: \$75,000 Duration: 1 year

Grant Proposal Project

Jing Zhao, Melissa Nuber, Olivia Meehan Pratt Institute, INFO-670: Linked Open Data, Spring 2022 Professor Christina Pattuelli

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Abstract

The Linked Cinema Project (LCP) seeks to enhance access to programming and collection information at the Anthology Film Archives through linked open data. This project will require restructuring existing data; converting the data to linkable, semantic triples; making the data publicly available; and integrating the improved data into the search and discovery functionalities of the Anthology Film Archives website. This data has immense research and commercial value, but is currently underutilized and under-connected. The Linked Cinema Project ultimately aims to open the possibility of inter-connection with other film archives and independent cinemas, and to advocate for the discovery of lesser-known films, directors, and cultural moments.

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Narrative

The Anthology Film Archives

The Anthology Film Archives (<u>http://anthologyfilmarchives.org/</u>) is an archive and theater located in the East Village of Manhattan. It is an international center for the preservation, study, and exhibition of film and video, with a particular focus on independent, experimental, and avant-garde cinema. Anthology screens more than 900 programs annually, preserves an average of 25 films per year (with 900 works preserved to date), publishes books and DVDs, and hosts numerous scholars and researchers.

Anthology was founded in 1970, and over the course of its 50-year history, has developed a rich tradition of film curation and public programming. The *Essential Cinema* series, for example, highlights films from the archive's *Essential Cinema* Repertory–a collection of 110 programs (330 titles) assembled in 1970-75 by Anthology's Film Selection Committee. They also regularly host director retrospectives; premieres and special screenings; new filmmaker screenings; and thematically curated series such as "Imageless Films" and "Folk Horror."

Anthology is and has long been an advocate for avant-garde, experimental, and independent film. The films they collect, preserve, and screen are often obscure or even completely unknown. Transforming the archive's collection and programming data to Linked Open Data could empower the discovery of lesser-known films and filmmakers-to the ultimate benefit of film scholarship and the archive itself.

The Linked Cinema Project

The Linked Cinema Project (LCP), in direct partnership with the Anthology Film Archives, aims to convert a sample selection of the archive's existing film datasets to linked data and open those linked datasets to public use via Wikibase. Future stages of the project involve implementing the web infrastructure needed to integrate the linked datasets with the Anthology website, and potentially partnering with other film institutions to create an interconnected research resource.

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Proposal Goal

LCP seeks a planning grant of \$75,000 to support 1) the analysis and cleanup of a sample dataset from Anthology, 2) data modeling, 3) the creation of RDF triples, and 4) the creation of initial access points to the linked data via Wikibase.

Statement of Need

On the Anthology Film Archives website, the program information pages are functionally dead-ends. Users cannot search for films based on directors, dates, genres, etc. This limits the discoverability of related films, materials, and events.

To improve accessibility to their collections and programming information, Anthology needs to ensure the consistency of data description; enable the interlinking of the archive's own data; improve data interoperability with other data publishers; and make data more discoverable. The Linked Cinema Project aims to provide solutions to those needs using the latest Linked Open Data standards.

Environmental Scan

There are a number of existing Linked Open Data projects from cultural heritage organizations that demonstrate the successful implementation of key project milestones, including: making data models and designating key conceptual entities; improving frontend features and functionality; partnering with other institutions; and utilizing Wikibase. The following environmental scan represents some of the projects most relevant to particular goals of LCP.

Linked Open Data at Carnegie Hall (https://data.carnegiehall.org/datalab/)¹

Carnegie Hall (CH) is a center of cultural and political expression, presenting and providing a venue for many different types of music and culture across multiple performance spaces. The Carnegie Hall archives database, the Orchestra Planning and Administration System (OPAS), is used to track every event that has occurred in the public performance spaces of Carnegie Hall². Starting in 2013, CH began publishing some of these records to its online Performance History Search³. The search has records for more than 49,000 events from 1891 to the present.

¹ Hudson, Rob, "Linked Open Data at Carnegie Hall. February 8, 2022," accessed May 10, 2022,

 $https://docs.google.com/presentation/d/181WCQ3Rn2eQajGUXza_vVPx8lYKeEI-M-9QFTTYjNbo/edit.$

² Carnegie Hall, "linked-data," *GitHub*, accessed April 27, 2022, https://github.com/CarnegieHall/linked-data.

³ https://www.carnegiehall.org/About/History/Performance-History-Search?q=&dex=prod_PHS

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

The performance data is structured around the concept of events. Each event entity has its own uniform resource identifier (URI) and includes metadata related to: date/time; venue; title; and participants, like conductor and/or an orchestra. Specific components of an event, such as works performed and performers, are also entities that have their own URIs. Anthology's programming, events, and screenings could be modeled in a similar way. We can learn from Carnegie Hall's use of entities and sub-entities to represent specific program components.

The Harvard Film Archive Project

(https://m.wikidata.org/wiki/Wikidata:WikiProject_Linked_Data_for_Production/Harvard_Fil m_Archive_Project)⁴

The Harvard Film Archive (HFA) is a division of Harvard Libraries dedicated to the collection, preservation, and exhibition of films. The HFA cinematheque, located on the Harvard University campus, is renowned for its diverse programming of films that are otherwise not screened regularly due to age, rarity, challenging content, or format⁵.

HFA's film data is browsable on its website⁶ and the archive is conducting an ongoing project to publish the data on Wikidata. On the Wikidata project page, there is a list of essential properties, suggested identifiers, and optional properties describing film directors, producers, screenwriters, films, and others. This is useful for us since our project would work with similar entities (programs, directors, and films). We can therefore refer to the Harvard Film Archive Project to learn about existing film description standards to improve our own data model.

⁵ Harvard Film Archive, "Cinematheque," accessed April 27, 2022, https://harvardfilmarchive.org/cinematheque.

⁴ Wikidata, "WikiProject Linked Data for Production/Harvard Film Archive Project," accessed April 27, 2022, https://m.wikidata.org/wiki/Wikidata:WikiProject_Linked_Data_for_Production/Harvard_Film_Archive_Project# References.

⁶https://harvardfilmarchive.org/collections/browse

Linked Open Data at the Anthology Film Archives

<u>ABSTRACT</u> <u>NARRATIVE</u> <u>ENVIRONMENTAL SCAN</u> <u>PROJECT DESIGN</u> <u>WORK PLAN</u> <u>APPENDIX</u> *Cinema Context* (<u>https://www.cinemacontext.nl</u>)⁷

Cinema Context (CC) is an online encyclopedia/database of Dutch Film culture, with information about more than 100,000 screenings in Dutch theaters, persons involved, and theater locations. The eventual objective of the project is to convert CC data to Linked Open Data.

The CC Data Model is founded in 5 interconnected entity types: persons, companies, venues, films and screenings. The *schema.org* vocabulary was used and Dublin Core was supplemented in certain instances where *schema.org* fell short, such as expressing a film's length and extent. CC's use of controlled vocabularies and adherence to the Resource Description Framework (RDF) demonstrates a prioritization of data interoperability that we should try to emulate with AFA.

The EUscreen Project (https://www.museumsandtheweb.com)⁸

The *EUscreen* project represents the European television archives and acts as a domain aggregator for Europeana, Europe's digital library. The main motivation for it is to provide unified access to a representative collection of television programs, secondary sources, and articles in order to allow students, scholars, the general public, and cultural heritage organizations to study the history of European television in its wider context.

The project's front-end, the *EUscreen* portal⁹, was designed after consulting the four primary user groups in order to define user requirements. The portal includes advanced features for specific use cases without overwhelming the users with complex interfaces. The Helsinki University of Arts and Design adapted a component-based conceptual model that accommodates this requirement. The portal showcases highlighted audiovisual items categorized by themes, news channels, and series, accompanying a keyword search bar on the top which enables users to filter the search result by language, decade, topic, provider, genre, country of production, and publisher.

⁷ Cinema Context, "About," accessed April 27, 2022.

https://www.cinemacontext.nl/cgi/b/bib/bib-idx?c=cccfilm;sid=a50434085deeae038e0e7d5f58f9c653;tpl=about.tpl;lang=en.

⁸ Museum and the Web, "Linking Europe's Television Heritage," last modified April 14, 2012,

https://www.museumsandtheweb.com/mw2012/papers/linking_europe_s_television_heritage.

⁹ https://euscreen.eu/

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

This is a good example of involving users from target/potential user communities to explicitly identify their needs. EUscreen also demonstrates a robust search and discovery system.

Additionally, *EUscreen* implemented a "server-less" front-end API where a proxy system handles the communication with the back-end services. *EUscreen*'s front-end system can be "installed" on any plain HTML web server without any need for server-side technologies. This means it can be hosted and moved to any location or multiple locations. It also means partners can use these APIs to integrate parts of the functionality in their own intranet and internet systems. This also could be a helpful reference for the potential development of a multi-institutional film research portal as a future stage of the Linked Cinema Project.

Project Design

The Anthology Film Archive currently uses a screening calendar as the primary search and navigation tool on its website (<u>http://anthologyfilmarchives.org/film_screenings/calendar</u>). The calendar can be filtered by labels such as "Series" and "Retrospectives." The content that shows up on the calendar is clickable and opens a list containing information about the screenings, including, if applicable, links to the particular program they are associated with. The archive's programming is a key conceptual element for organizing the film screenings within the structure of the website. There are five program categories as indicated by the website's navigation: Essential Cinema, Premieres/Revivals, Retrospectives, Series, Special Screenings, AFA preservations, and New Filmmakers. There is an existing "series" subdomain and there are individual ID numbers for each existing program:

BRUCE & NORMAN YONEMOTO (RETROSPECTIVE)

http://anthologyfilmarchives.org/film_screenings/series/54353

IMAGELESS FILMS, PART 1 (SERIES)

http://anthologyfilmarchives.org/film_screenings/series/54404

The primary goal of LCP is to create similar linkable URIs and ID numbers for entities beyond just screenings and programs, including films, directors, locations, etc. This linked data will be openly available for querying, visualization, or any other form of use and reuse.

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

The Linked Cinema Project is seeking an initial planning grant to develop an effective model for converting AFA's film collection into linked open data over the course of a year long project with four distinct, consecutive stages:

- 1. Analyzing AFA's collection data and selecting a sample dataset
- 2. Data modeling
- 3. Extracting entities and converting to RDF
- 4. Creating access points, testing, and disseminating results

Our initial goal is to test the effectiveness of our data model using a random sample of 90 dissimilar film items from AFA's repository.

We intend to use Wikibase to host this project, which will allow us to use our own data model; manage our dataset openly; create our own entities and values; and channel our dataset to Wikidata. Our work plan, therefore, is based heavily around utilizing the features of Wikibase.

All results and output from this year-long project will be made freely and publicly available online. We will publish our results in open repositories so that our target community, consisting of film archives and similar GLAM institutions across the world can use it as a model for converting and implementing linked open data in their own institutions. The ultimate, eventual goal of The Linked Cinema Project is to facilitate wider discovery of non-traditional film, histories and resources by promoting valuable linkages across institutional datasets. Future phases of this project beyond our work with Anthology internally, will focus on external, user-facing development and partnerships with other institutions.

Project Budget

We are seeking \$75,000: \$60,000 will be allocated to cover project personnel, with the remaining \$15,000 to cover equipment, technology and any logistical needs.

Project Personnel

The Linked Cinema Project team will consist of two part-time consultants: One data scientist with linked open data experience, and one Anthology Film Archives archivist or subject specialist.

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX Work Plan

Phase 1: Analyzing Anthology Film Archives and Selecting Sample Dataset

Proposed Timeline: 2 months

The first stage of The Linked Cinema Project is to do a general inventory of the anthology collections, alongside an AFA archivist/project consultant. This will include an inventory of film collections, film materials, and all related metadata records detailing the contents of the Anthology's collection. This inventory will give us a better understanding of decisions that have been made by AFA around information architecture and how archival data is structured not only the physical archive but on AFAs website and digital archive as well.

After we have a well-rounded sense of the archive's collections, we will select 90 samples from AFA repositories that represent our primary entities: program, film and director. The sample will include a random selection of 30 program descriptions and 30 film notes already available online, along with 30 paper resources based on individual directors. The digital resources will then be converted into Microsoft Word documents and the paper materials will be scanned and converted to plain text using Pomodoro.¹⁰

Phase 2: Data Modeling

Proposed Timeline- 2 months

In Phase 2, our first step will be to find a formal standard for developing sustainable data models. Appendix A, B, and C show preliminary data model diagrams for the three primary entities of TLC (program entity, film entity, director entity). We will complete the existing preliminary data model by including additional sub-entities and literals that fully describe the primary entities. We used Wikidata properties in the preliminary data model, but when warranted, predicates from other relevant ontologies, such as linked.art¹¹, CIDOC CRM¹², and schema.org¹³ could also be used to enable the full and accurate description of primary and secondary entities.

¹⁰ https://pomodoro.semlab.io/

¹¹ https://linked.art/

¹² https://www.cidoc-crm.org/

¹³ https://schema.org/

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Our next step would be to choose controlled vocabularies. We will look for well-established standards to ensure consistent data description and to provide reusable identities for shared values across datasets. We could potentially use the Union List of Artist Names (ULAN) to control the names of people; the Getty Thesaurus of Geographic Names (TGN) to control the names of places; and Library of Congress Subject Headings or the Getty Art and Architecture Thesaurus (AAT) to control subject and genre terms. When needed, we could use domain specific vocabularies or we could create our own values on Wikibase.

Lastly, we will review the completed preliminary data model with selected representatives from our three primary user groups (scholars, the general public, and other film organizations). The data model will be revised based on the results of those discussions.

Phase 3: Extracting entities and creating RDF triples

Proposed Timeline- 4 months

In this phase, we will import the 90 Word Documents generated from our data sample into SÉLAVY¹⁴ to clean up text manually, and 'block' them into chunks. Afterward, we will send the text blocks to the Name Entity Recognition (NER)¹⁵ server–a tool for recognizing the name entities in text files. To ensure accuracy, we will inspect the text and manually select any names that were not recognized by the tool. Using Wikibase, we will assign an entity type (program, film, director, organizer, place, etc.) and an URI to each of the names recognized.

Next, we will create RDF triples for recognized name entities by using the 'Export Rules' function in the NER tool. The following is an example of an RDF triple:

Subject: City Symphonies (http://base.anthologylab.io/wiki/Item:Q50342)
↓
Predicate: has part or parts (https://www.wikidata.org/wiki/Property:P527)
↓
Object: Man with a Movie Camera (https://base.anthologylab.io/wiki/Item:Q82925)

Once the triples are created, we will export them to Wikibase.

¹⁴ A document to data tool developed by the Semantic Lab at Pratt, http://159.89.242.202:3000/

¹⁵ https://semlab.io/DADAlytics_ner_demo/

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Phase 4: Creating Access Points and Disseminating Data

Proposed Timeline: 4 months

In the final stage of our planning project, we will focus on creating and maintaining points of access to our dataset. Through Wikibase, we can allow the export of our data as RDF, JSON and XML and we will also publish our data on Github to further encourage access and collaboration. SPARQL queries of our data can also be performed via Wikibase.

Furthermore, we will propose a new Wikidata Property ("Anthology Film Archives ID"). We will draft and submit our proposal through the Wikidata discussion board. To prepare this proposal, we will identify film archivists, GLAM workers and data scientists who understand the need for a new property related to avant-garde/experimental film and invite them to contribute to the Wikidata discussion around our proposed property. This outreach will be important given that Wikidata properties can only be created by Wikidata administrators after a period of discussion that yields support for the property being proposed and little to no opponents¹⁶.

Finally, we will distribute the results of our year-long effort, publishing our findings in open access repositories and applying to present at related information shows/symposiums such as LD4. We will create a short video about our project to present at conferences and to circulate online.

Future Work

In the future, the Linked Cinema Project will shift its focus from internal development at AFA to external projects and partnerships with similar institutions. Next steps for this project include:

- □ Redesigning the front end of AFA's website to increase visibility of collections and improve search function.
- Researching target institutions and creating outreach and educational plans to encourage collaboration and linking to external datasets.

¹⁶ https://www.wikidata.org/wiki/Wikidata:Property_proposal

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

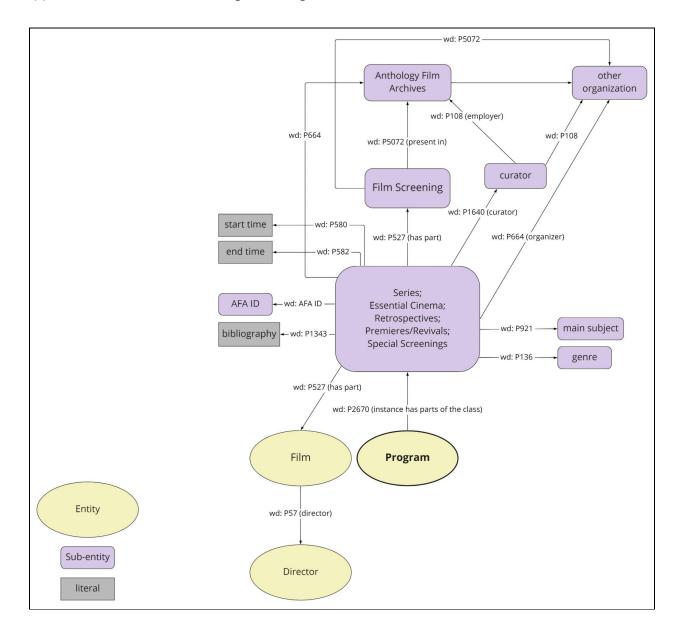
- Creating a task force to address archival/non-traditional film LOD standards.
- □ Planning an LOD symposium and inviting target institutions and partners to attend and participate.
- Designing a Linked Cinema Project website that aggregates linked archival materials from all partner institutions.
- Developing best practices for handling structured data related to nontraditional films including films with no known release dates, films of unique formats or durations, and orphan films where copyright holders are unknown.

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Appendix

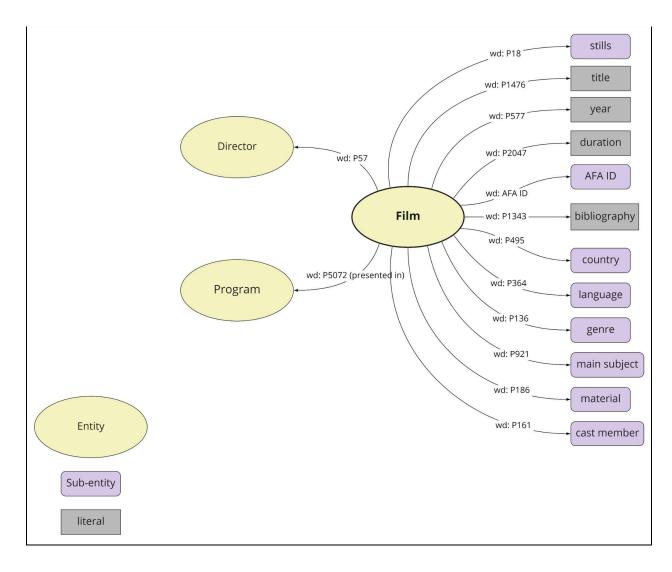
Appendix A. AFA data model diagram: Program



Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

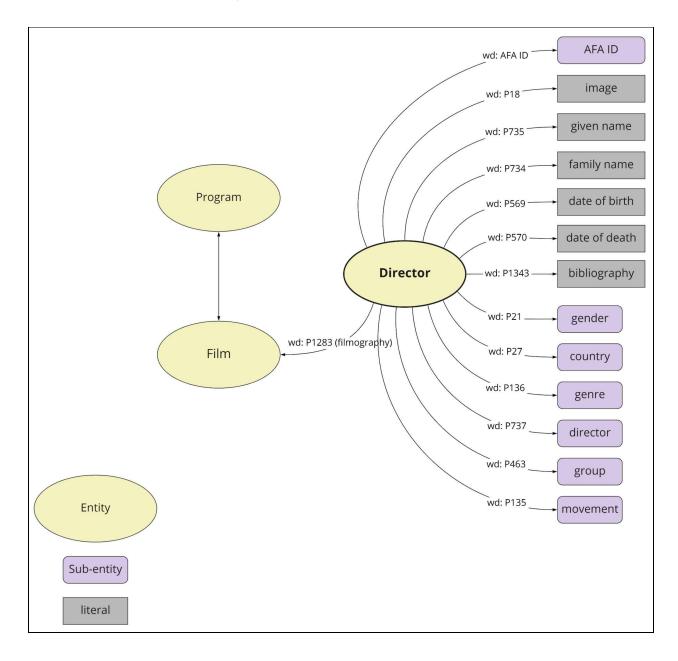
Appendix B. AFA data model diagram: Film



Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Appendix C. AFA data model diagram: Director



Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Appendix D. AFA Data Model table: Program

Subject	Predicate	Object	Object is an entity?	Literal type
Program	instance has parts of the class (wd: P2670)	Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	has part or parts (wd: P527)	Film Screening	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	organizer (wd: P664)	Anthology Film Archives, other organization	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	curator (wd: P1640)	curator	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	has part or parts (wd: P527)	film	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	start time (wd: P580)	start time	no	xsd:dateTime
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	end time (wd: P582)	end time	no	xsd:dateTime
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	wd: Anthology Film Archives ID	Anthology Film Archives ID	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	described by source (wd: P1343)	bibilography	no	string
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	main subject (wd: P921)	main subject	yes	URL
Series, Essential Cinema, Retrospectives, Premieres/Revivals, Special Screenings	genre (wd: P136)	film genre	yes	URL

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Appendix E. AFA Data Model table: Film

Subject	Predicate	Object	Object is an entity?	Literal type
film	director (wd: P57)	director of the film	yes	URL
film	presented in (wd: P5072)	program that the film presented in	yes	URL
film	image (wd: P18)	stills of the film	yes	URL
film	title (wd: P1476)	title	no	string
film	publication date (wd: P577)	year or more specific data	no	xsd:dateTime
film	duration (wd: P2047)	duration	no	string
film	wd: Anthology Film Archives ID	Anthology Film Archives ID	yes	URL
film	described by source (wd: P1343)	bibliography	no	string
film	country of origin (wd: P495)	country	yes	URL
film	original language of film or TV show (wd: P364)	language	yes	URL
film	genre (wd: P136)	film genre	yes	URL
film	main subject (wd: P921)	main subject	yes	URL
film	made from materiel (wd: P186)	material or format of the film	yes	URL
film	cast member (wd: P161)	cast member	yes	URL

Appendix F. AFA Data Model table: Director

Subject	Predicate	Object	Object is an entity?	Literal type
director	filmography (wd: P1283)	list of films person has directed	yes	URL
director	wd: Anthology Film Archives ID	Anthology Film Archives ID	yes	URL
director	image (wd: 18)	photograph or other depiction of person	no	string
director	given name (wd: P735)	first name or another given name	no	string
director	family name (wd: P734)	surname	no	string
director	date of birth (wd: P569)	year or more specific date	no	xsd:dateTime
director	date of death (wd: P570)	year or more specific date	no	xsd:dateTime
director	described by source (wd: P1343)	bibliography	no	string
director	gender (wd: P21)	male, female, intersex, transgender female, transgender male, non-binary, or other values that reflect an individual's self-identified gender	yes	URL
director	country of citizenship (wd: P27)	country	yes	URL
director	genre (wd: P136)	film genre assciated with this person	yes	URL
director	influenced by (wd: P737)	a person that influenced this person's films	yes	URL
director	group (wd: P463)	the film group the person was in	yes	URL
director	movement (wd: P135)	the movement the person was in	yes	URL

Linked Open Data at the Anthology Film Archives

ABSTRACT NARRATIVE ENVIRONMENTAL SCAN PROJECT DESIGN WORK PLAN APPENDIX

Appendix G. References

Carnegie Hall. "linked-data." *GitHub*. Accessed April 27, 2022. https://github.com/CarnegieHall/linked-data.

Cinema Context. "About." Accessed April 27, 2022. https://www.cinemacontext.nl/cgi/b/bib/bib-idx?c=cccfilm;sid=a50434085deeae038e0e7d5f5 8f9c653;tpl=about.tpl;lang=en.

Harvard Film Archive. "Cinematheque." Accessed April 27, 2022. https://harvardfilmarchive.org/cinematheque.

Hudson, Rob. "Linked Open Data at Carnegie Hall." February 8, 2022. https://docs.google.com/presentation/d/181WCQ3Rn2eQajGUXza_vVPx8lYKeEl-M-9QFTTYjNb o/edit.

Museum and the Web. "Linking Europe's Television Heritage." Last modified April 14, 2012. https://www.museumsandtheweb.com/mw2012/papers/linking_europe_s_television_herita ge.

Wikidata. "WikiProject Linked Data for Production/Harvard Film Archive Project." Accessed April 27, 2022.

https://m.wikidata.org/wiki/Wikidata:WikiProject_Linked_Data_for_Production/Harvard_Film_Archive_Project#References.