Cinema Context as Linked Open Data*

Converting an online Dutch film culture dataset to RDF

Leon van Wissen[1][0000−0001−8672−025X], Thunnis van Oort[2][0000−0001−8912−0508], Julia Noordegraaf[1][0000−0003−0146−642X], and Ivan Kisjes[1]

[1] University of Amsterdam, The Netherlands
{l.vanwissen, j.j.noordegraaf, i.kisjes}@uva.nl
[2] Radboud University Nijmegen, The Netherlands
thunnis.vanoort@ru.nl

Abstract. This paper describes the process of converting Cinema Context, an online dataset on Dutch film culture, into Linked Open Data. It covers our experiences in this conversion process from the steps of data cleaning and modeling, up to publishing and evaluating the result through a case study.

Keywords: Cinema History · Digital Humanities · Linked Open Data

1 Introduction

Cinema Context (CC) is an online encyclopedia on Dutch film culture since 1896 [1]. Built on top of a MySQL database the website www.cinemacontext.nl offers both an informational view as well as a research environment on places, persons and companies involved in more than 100k film screenings in the Netherlands. The website allows a visitor to search and extract the data, though this is limited to the offered capabilities of the faceted search. To provide access to the full dataset and to boost its interactivity, we have now published it as RDF.

In this paper, we describe the process of converting a relational database with Cultural Heritage (CH) data into Linked Open Data (LOD). It also gives an example of the potential that this format offers. Specifically, converting this dataset to LOD brings opportunities for broadening and renewing historical and cultural research by allowing more flexible linking to other (linked) datasets on for instance buildings, persons, heritage objects, and locations. Researchers in the Digital Humanities (DH) and CH communities have indicated a need [5] to be able to query CC in connection with external data via e.g. a SPARQL endpoint, in order to research the role of cultural and socio-economic status in processes of cultural consumption. Moreover, the selection of appropriate vocabularies and thesauri required close collaboration between data specialists and domain experts and has functioned as de facto training in working with RDF and the SPARQL query language for scholars working in DH.

* This project was partly financed by a DANS Small Data Project that stimulates projects that adhere to the FAIR guiding principles for scientific data management.

2 Model

The CC data model [6] has its foundation in five interconnected entity types: persons, companies, venues, films, and screenings. These core entities are given a persistent identifier which serves as URI. Due to its ease of use and its increasing applicability in the DH and CH the SCHEMA.ORG vocabulary is particularly suitable for modeling the contents of CC. Although the database is not aimed at describing present-day or future events, the individual entities in CC contain sufficient information to model the classes and properties in this vocabulary. How we model these entities is described below (see Fig. 1).

2.1 Entity types

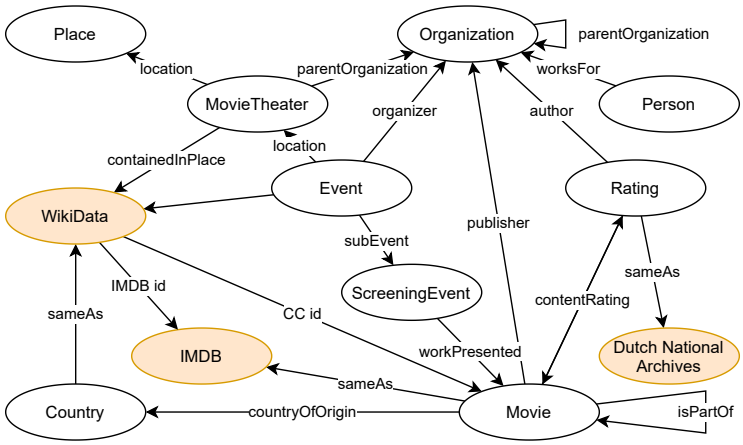

Fig. 1. Classes and their interrelation in the CC RDF data. The white nodes are classes from the SCHEMA.ORG vocabulary.

Films and ratings We model each film that is described in CC as a `schema:Movie` and provide information on its name, alternate name, production year, country of origin, format, and extent. A `schema:sameAs` property is available for every film to refer to its entry on www.imdb.com. Ratings for these films are modeled as `schema:Rating` and `schema:CreativeWork` and come from the archive of Film Screening Reports 1928-1960 which is held by the Dutch National Archives. The value in the `schema:sameAs` property of a rating points to the respective index in this collection.

Persons Persons in the CC data can be owners or employees of theaters and companies. They are modeled as `schema:Person`, provided with biographical information such as birth dates if available. A person's name is modeled through the Person Name Vocabulary (`pnv`)[3].

[3] https://w3id.org/pnv#

Venues and companies Theaters or venues are organizations situated at a specific location (`schema:Place`, including geometry) and with a specific name. Usually, they are owned by a company or person. Each theater is a `schema:MovieTheater`, or a `schema:EventVenue` for general-purpose venues. If available, information on the seating capacity, number of screens, and lifespan is also given. The companies in this dataset are organizations that run cinemas or distribution companies and are modeled as an instance of `schema:Organization`.

Events and screenings We distinguish two event classes: `schema:Event` for cinema programs, and `sem:Event` from the Simple Event Model (`sem`)[4] for generic events, such as a venue's construction history. Instances of the first consist of one or more `schema:subEvent`s of type `schema:ScreeningEvent` (filmic) or `schema:TheaterEvent` (non-filmic), held in a specific theater on one or more dates, using a particular program name. A `schema:startDate` indicates the day a program started: normally, a cinema program would be screened for a week. Deviations from this norm are usually annotated as `schema:description`.

2.2 Other and qualifiers

The SCHEMA.ORG vocabulary is complemented where it falls short, for instance when describing person names, legal entity types, and special cinema types (e.g. traveling cinema), but also when expressing a film's length and extent, for which we use properties from Dublin Core[5]. Additionally, if a date is given in a less precise format than `xsd:date`, then the `sem` time stamp properties have been used to supply a proper date value. For consistency and usability, and to indicate a temporal restriction or (un)certainty, the `sem` properties are always present, even if an exact date is given.

This is also the case when resources are temporally restricted. We use the `schema:Role` class to express a specific time frame in which a certain property value relation is valid. This class can be used in any object position and extends the triple with the same property, whilst incorporating additional information as qualifier, such as specific roles some entity played, or a start and end date. This way of modeling is used consistently in the data to boost its queryability, even when there is nothing to qualify. A description of less prominent auxiliary classes can be found in the dataset's documentation pages (see Section 3).

3 Documentation and code

Documentation pages [3] were built to accompany the constructed RDF and include an explanation of the used vocabulary, modeling and SPARQL query examples, reports on hands-on sessions, and general information about the project.

[4] http://semanticweb.cs.vu.nl/2009/11/sem/
[5] http://purl.org/dc/terms/

Both the documentation and the code that converts the MySQL database are available in a git repository [3]. A pipeline is built in such a way that new LOD can be generated instantly whenever a new dump of the database is made. The latest dump of the MySQL dataset can be found at DANS [2].

4 Case Study: International Orientation Index

A case study[6] serves to illustrate the potential of connecting the CC dataset with other knowledge graphs. It replicates the analysis of economic film historian Peter Miskell et al. [4] and their 'international orientation index'. Miskell et al. propose this index to investigate the relative success of Hollywood productions abroad in the post-war reconstruction period and state that American productions with a high proportion of non-American creative talent and content[7] have fared better at non-American box offices.

We can test this hypothesis for the Dutch film market by analyzing programming data from CC. What is missing in our dataset is information on box office revenues, but this value can be approximated by the number of screening weeks under the assumption that a film with more screenings generates higher revenue. To approximate the variables Miskell et al. used to construct their index, we can apply the information available for films in Wikidata. Instead of assigning a 0, 1 or 2 score to a criterion, we assigned a relative score (0.0-1.0) to the variables, indicating the extent of 'internationalisation' (or rather: 'non-Americanness') in a category. For a total of 8,836 films (5,495 Hollywood productions), we gathered information on six categories through a SPARQL query, each retrieved with a particular Wikidata property path (e.g. the film's director, followed by his/her country of citizenship). Examples of this calculation are shown in Table 1.

Calculating a correlation coefficient between the number of screenings and the relative internationalness of the Hollywood produced part of our corpus by using Pearson's r indicates that there is a very weak correlation of 0.130 when we consider films (N=3418) for which we have information in at least three categories and 0.137 when we consider films (N=1340) for which we have at least four categories (below three is not sufficiently representing internationalness; over five reduces the corpus size too much). Though positive, and thus indicating that internationally-oriented films in the Netherlands perform slightly better than fully American ones, we should further refine this proof of concept in future studies before making solid claims.

5 Summary

This project shows that the SCHEMA.ORG vocabulary can easily be applied to cultural heritage data and deemed fit for modeling our (research) dataset. With

[6] A more detailed explanation of this and other case studies, including code and data, can be found in the documentation pages [3] under 'events'.

[7] Measured based on (1) nationality of leading actors, directors, screenwriters, (2) setting, and (3) national provenance of the source text.

Table 1. Individual examples of calculating this score. The relative total scores are calculated by dividing the total score over the number of available variables.

Category	***Anna Karenina (1935)***	***Casablanca (1942)***	***Key Largo (1948)***
CC id	F001809	F020802	F015663
Wikidata id	Q561208	Q132689	Q830773
Screenings	37	28	No data available
Director	0.0	0.50	0.0
Screenwriter	0.5	0.0	0.0
Cast	0.53	0.58	0.125
Narrative	1.0	1.0	0.0
Shooting	No data available	0.0	No data available
Source author	1.0	0.0	0.0
Total (relative)	**3.03 (0.61)**	**2.08 (0.35)**	**0.125 (0.03)**
Miskell et al.	12	7	1

some small additions, we were able to capture and publish this dataset in LOD, and thereby make it more readily available for (re)usage in the DH and Social Sciences. The case study demonstrates how such a dataset can be operationalized in the workflow of a DH research project. For the time being, the LOD version of CC exists besides the original database and accompanying website, but ideally, these will be merged and/or further integrated in a future version.

Acknowledgements

The project was a collaboration between the CC editorial staff, Library UvA, and Menno den Engelse (Islands of Meaning).

References

1. Dibbets, K.: Cinema Context and the genes of film history. New Review of Film and Television Studies **8**(3), 331–342 (2010)
2. Dibbets, K.: Cinema Context. film in Nederland vanaf 1896: Een encylopedie van de filmcultuur (2018). https://doi.org/10.17026/dans-z9y-c5g6
3. den Engelse, M., van Wissen, L., van Oort, T., Noordegraaf, J.: Cinema Context in RDF (2020). https://doi.org/10.17026/dans-z64-mrvb, https://uvacreate.gitlab.io/cinema-context/cinema-context-rdf/
4. Miskell, P., Li, Y.: Hollywood studios, independent producers and international markets: Globalisation and the US film industry c. 1950–1965. Henley Business School (2014)
5. Noordegraaf, J., et al.: Semantic deep mapping in the Amsterdam Time Machine: Viewing late 19th- and early 20th-century theatre and cinema culture through the lens of language use and socio-economic status. CCIS (2021 forthcoming)
6. van Oort, T., Noordegraaf, J.: The Cinema Context database on film exhibition and distribution in the Netherlands: A critical guide: Arts and media. RDJ for the SSH **5**(2), 91–108 (2020). https://doi.org/10.1163/24523666-00502008