



Film Noir - exploratory analysis



Horizon Europe Data Management Plan

23 January 2023

*Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.4.4 (dsw:root:2.4.4).*

HISTORY OF CHANGES		
Version	Publication date	Changes
<i>There are no named versions</i>		

Contributors

The following contributors are related to the project of this DMP:

- **Catalina Cruceanu**
c.cruceanu@student.rug.nl
Roles: *Contact Person, Data Collector, Data Curator, Editor*
Affiliation: [*University of Groningen*](#)
- **Ece Demitras**
e.demirtas@student.rug.nl
Roles: *Data Curator, Editor*
Affiliation: [*University of Groningen*](#)
- **Shannon Wietsma**
s.k.wietsma@student.rug.nl
Roles: *Data Curator, Data Manager, Editor, Project Leader*
Affiliation: [*University of Groningen*](#)
- **Z. Wang**
z.wang.75@student.rug.nl
Roles: *Data Manager, Editor*
Affiliation: [*University of Groningen*](#)

Projects

We will be working on the following project and for those are the data and work described in this DMP.

Film Noir - an exploratory analysis

Acronym: *FNEA*

Start date: *2022-12-20*

End date: *2023-01-24*

Funding: [*Deutschen Filmförderfonds*](#): *grant number not yet given (planned)*

‘Film noir’ is a term you’ve probably heard hundreds of times, and has long been regarded as one of cinema’s most intriguing styles that defined post-war American cinematography. The genre developed iconic cinematic motifs and tropes that inspire filmmakers today. For this reason, our group decided to focus our research project on the topic of Film- Noir movies, and thereby we searched online for a suitable dataset. We will use the dataset “Film Noir: They Shot Dark Pictures, Didn’t They?” (<https://www.kaggle.com/datasets/kabhishm/film-noir-they-shot-dark-pictures-didnt-they>).

1. Data Summary

We will need to harmonize different sources of existing data.

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values**

It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **Film Noir Dataset** (published)

The dataset has the following identifiers:

- URL: <https://www.kaggle.com/datasets/kabhishm/film-noir-they-shot-dark-pictures-didnt-they>

We will distribute the dataset using:

- *Domain-specific repository*: Github. We don't need to contact the repository because it is a routine for us.

There will be different versions of this data over time; the versions will be dated.

We will not be adding a reference to any data catalogue because the data will be stored in a repository that is the prime source of data for re-use in the field.

There are the following 'Minimal Metadata About ...' (MIA...) standards for our experiments:

-
-

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

We made a SOP (Standard Operating Procedure) for file naming. Every member of the project will be naming files according to the file's content. . We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open immediately.

Limited embargo will not be used as all data will be opened immediately.

Metadata will be openly available without instructions how to get access to the data.

Metadata will be available in a form that can be harvested and indexed (managed by the used repository / repositories).

Our data is legally not copyrightable, there is no legal owner.

For our produced data, conditions are as follows:

- **Film Noir Dataset** (published)

The distributions will be accessible through:

- *Domain-specific repository*: Github. We don't need to contact the repository because it is a routine for us.

A user of this data can use it without any specific software.

The dataset will be published when the project is wrapped up.

2.3. Making data interoperable

We will be using the following data formats and types:

- **Comma-separated Values**

It is a standardized format.

2.4. Increase data re-use

The metadata for our produced data will be kept as follows:

- **Film Noir Dataset** (published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.

As stated already in Section 2.2, all of our data can become completely open immediately.

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will be instrumenting the tools into pipelines and workflows using automated tools.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. The management and planning of other research outputs is done separately and is included as appendix to this DMP. Still, we benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

We will be archiving data (using so-called 'cold storage') for long term preservation after the project but also already during the project. The costs for archiving data will be paid from the grant. The minimum lifetime of the archive is 15 years. The archival period can be extended – library or archive staff will decide. The decision whether or not to extend the renewal be based on the actual use of the archived data. Data formats of data in cold storage will be upgraded if they become obsolete. Archived data will be migrated regularly to more modern storage media (e.g. newer tapes).

None of the used repositories charge for their services.

Catalina Cruceanu is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Catalina Cruceanu , Ece Demitras, and Shannon Wietsma are responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Catalina Cruceanu is responsible for finding, gathering, and collecting data.

Shannon Wietsma and Z. Wang are responsible for maintaining the finished resource.

To execute the DMP, additional specialist expertise is required and we have such trained support staff available.

We do not require any hardware or software in addition to what is usually available in the institute.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (<https://...>). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

The archive will be stored in a remote location to protect the data against disasters. The archive need to be protected against loss or theft. It is clear who has physical access to the archives.

6. Ethics

For the data we produce, the ethical aspects are as follows:

- **Film Noir Dataset**
 - It does not contain personal data.
 - It does not contain sensitive data.

Data we collect

We will not collect any data connected to a person, i.e. "personal data".

7. Other issues

We use the [Data Stewardship Wizard](https://researchers.ds-wizard.org) with its *Common DSW Knowledge Model* (ID: dsw:root:2.4.4) knowledge model to make our DMP. More specifically, we use the <https://researchers.ds-wizard.org> DSW instance where the project has direct URL: <https://researchers.ds-wizard.org/projects/17d66e89-d256-4219-9726-95f10efea8f8>.

We will not be using any extra national, funder, sectorial, nor departmental policies or procedures for data management.