CINEMA

Cinema Database Specification
Dietrich Release v1.0

Document Version 1.1
(11 October 2017)
LA-UR-17-25072

by

Jon Woodring woodring@lanl.gov
David Rogers dhr@lanl.gov
James Ahrens ahrens@lanl.gov
John Patchett patchett@lanl.gov

Los Alamos National Laboratory
Bikini Atoll Rd., SM 30
Los Alamos, NM 87545
cinema@lanl.gov

Contents

1 Introduction

This document is a specification for Cinema databases. It is specification D (Dietrich), version 1.0. This is a different approach than past specifications [1] [2], but is complimentary in spirit and function to them. This version is a simple embodiment of the new approach, which can be easily adapted to a wide range of use cases, and is designed for quick adoption by scientists, programmers and others.

See the Cinema website (`http://www.cinemascience.org`) for additional information, and contact the cinema community (`cinema@lanl.gov`) or the authors of this document with questions.

2 Cinema Overview

Extreme scale scientific simulations are leading a charge to exascale computation, and data analytics runs the risk of being a bottleneck to scientific discovery. Due to power and I/O constraints, we expect in situ visualization and analysis will be a critical component of these workflows.

Options for extreme scale data analysis are often presented as a stark contrast: write large files to disk for interactive, exploratory analysis, or perform in situ analysis to save detailed data about phenomena that a scientist knows about in advance. Cinema represents a novel framework for a third option a highly interactive, data artifact-based approach that promotes exploration of simulation results, and is easily accessed through database specifications. This approach supports interactive exploration of a wide range of results, while still significantly reducing data movement and storage.

More information about the overall design of Cinema is available in the paper *An Image-based Approach to Extreme Scale In Situ Visualization and Analysis* [3].

A Cinema Database supports the following three use cases. Taken together, these support a novel method for interactively exploring artifacts from extremely large datasets.

1. Searching/querying of meta-data and data artifacts. Samples can be searched purely on metadata, content, position, time, or a combination of all of these.

2. Interactive visualization of sets of data artifacts.

3. Playing interactive visualizations, allowing the user on/off control of elements in the visualization.

2.1 What is a Cinema Database?

A Cinema database is a set of precomputed data artifacts that can be queried and interactively viewed. The user can decide what types of components comprise the database, based on the type of interaction that is desired with the final database. A general design philosophy of Cinema is that applications reading and viewing a Cinema database can ignore data and determine which operations to perform. This promotes a wide range of possible interactions with the data - not just the ones imagined by the creator of the database.

Previous Cinema database specifications have concentrated on the notion that a database is a set of results sampled by visualization parameters. In this specification, we abstract this to include typical use cases from experiments by scientists.

A scientist often has a spreadsheet with data about parameters for an experiment. This results in a set of parameters that map to a particular result - a graph, a sensor image, or other image-based data. These are a natural abstraction from previous Cinema databases. Figure 1 shows this mapping of parameter sets to results. A collection of these mappings can be easily expressed in a cinema database, using this specification. This is detailed in later sections.

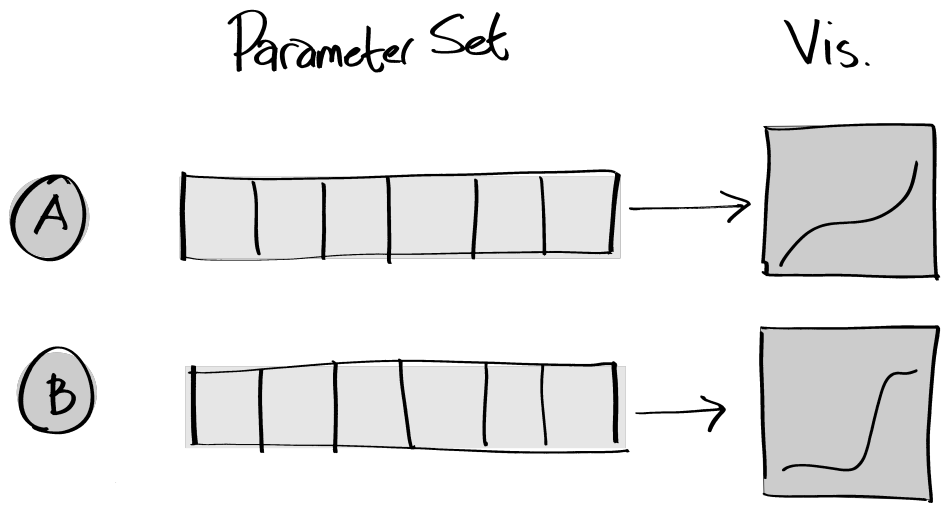

Figure 1: Diagram showing a typical set of data that a scientist has for an experiment or simulations. Some set of parameters (A or B) has been used to create a visualization - a graph, an image captured from a sensor, or other data. These parameter sets can include, for example, settings on an experimental machine, inputs to a simulation, or measurements taken by a sensor. Each one of the parameter sets thus defines a unique result. Taken together, a set of these parameter sets constitutes a database of results, and a scientist often tracks this database in a spreadsheet. These parameter set/image pairings form the basis for the simplest Spec D database.

3 The Cinema D (Dietrich) Specification

Cinema D (Dietrich) specification is a table-based data model for scientific data. In version 1.0 of this specification, we take the simplest approach to this new model, in order to simplify adoption and interaction with Cinema. Later versions of this specification will include additional capabilities, and will be detailed in separate documents.

3.1 Required elements

A Spec D version 1.0 Cinema database is directory that contains all of the data in the database. The directory must have the following files. Additional files or directories may be present, but they are ignored by Cinema.

- a folder named `database_name.cdb` containing:
  - `data.csv`, a comma-separated value file. This file is specified in section Section3.2. The presence of this named file exactly indicates that this is a Spec D, version 1.0 file, due to the fact that there is no other metadata in the database. In particular, *if this file is present, the database is by definition a Spec D, version 1.0 database.*
  - optional data files, referenced in the csv file
  - there may be additional files in the directory, but they are ignored by this specification.

3.2 The `data.csv` File

The main data file for this specification is a comma-separated-value file that has the following requirements. There are two types of csv files: **Data Only** and **Data with File References**, which are detailed in the following sections. Note that *new data can be added to the file, and existing data deleted, as long as the requirements are met.*

3.2.1 Type I: Data Only

A **Data Only** `csv` file is simply a table of float values. There is no external data referenced. The requirements on the file are:

1. The file is UTF-8 encoded.
2. The file follows the specification rfc4180 [4].
3. Each line is fully populated (there can be no missing values).
4. The first line of the file is **required**. It is unique, and is **strings** that are labels for the columns in the file.
5. All other lines in the file are data. Each value in these lines must be a **float**.

Example This type of `data.csv` file is self-contained.

```
timestep,time value,x,y,z
1.0,0.1,1.0,1.1,1.2
2.0,0.2,2.0,2.1,2.2
3.0,0.3,3.0,3.1,3.2
4.0,0.4,4.0,4.1,4.2
```

In this case, the cinema database would contain the following files:

```
database.cdb/
    data.csv
```

3.2.2 Type II: Data with file reference

A **Data with File Reference** `csv` file is a table of float values, with a final column that references an external file containing additional data. The file can be of any type or extension. The requirements on this type of `csv` file are:

1. The file is UTF-8 encoded.
2. Each line is fully populated (there can be no missing values).
3. The first line of the file is **required**. It is unique, and is **strings** that are labels for the columns in the file. The last label is *required* to be the keyword `FILE`.
4. All other lines in the file are data. In the data lines:
   (a) All values except the last value must be **floats**.
   (b) The last column is a **string** that is a POSIX file path relative to the base directory (the `name.cdb` path) of the Cinema database. This is the location of the data matched to the values in the columns for that row. The file can be of any type or extension.
      i. Cinema Spec D readers are only required to handle (display) image files of type JPG, JPEG, PNG, and GIF.

Example This `data.csv` example shows the `FILE` keyword on the last column, and POSIX string paths.

```
timestep,time value,x,y,z,FILE
1.0,0.1,1.0,1.1,1.2,img/001.jpg
2.0,0.2,2.0,2.1,2.2,img/002.jpg
3.0,0.3,3.0,3.1,3.2,data/003.jpg
4.0,0.4,4.0,4.1,4.2,data/004.png
```

In this case, the cinema database would contain the following files:

```
database.cdb/
    data.csv
    img/
        001.jpg
        002.jpg
    data/
        003.jpg
        004.png
```

4 Contacts and Further Information

For further information, email the cinema mailing list at `cinema@lanl.gov`, or contact the authors of this document. Additional information is available at the cinema website, `http://www.cinemascience.org`.

5 Acknowledgements

The image used on the cover page is a publicity photo of Marlene Dietrich for the film Shanghai Express (1932). This photo is in the public domain [5].

6 Bibliography

References

[1] D. Rogers, J. Ahrens, and J. Patchett, "Cinema simple database specification," Tech. Rep. LA-UR-15-20572, Los Alamos National Laboratory, January 2015.

[2] D. Rogers, J. Woodring, J. Patchett, D. DeMarle, and B. Geveci, "Cinema database specification chaplin release," Tech. Rep. LA-UR-15-20645, Los Alamos National Laboratory, January 2017.

[3] J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, and M. Petersen, "An image-based approach to extreme scale in situ visualization and analysis," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, (Piscataway, NJ, USA), pp. 424–434, IEEE Press, 2014.

[4] Y. Shafranovich, "Common Format and MIME Type for Comma-Separated Values (CSV) Files." RFC 4180, Oct. 2005.

[5] Deanlaw, "Marlene dietrich in shanghai express (1932) by don english," 2016. `https://commons.wikimedia.org/wiki/File:Marlene_Dietrich_in_Shanghai_Express_(1932)_by_Don_English.png`.