PIDs Go To The Movies

Raymond Drewry
MovieLabs/EIDR
rdrewry@movielabs.com

PIDapalooza 2019 – Dublin
What I wanted

• **Coverage of a wide range of works** from global sources (excluding user-generated content for now)

• **Appropriate granularity of identification** (covering the abstract concept of an underlying work and all its many variations)

• **Reliable, free access** to the identifiers and their metadata (i.e. the identifiers are resolvable, and anyone can use them)

• **Connection to other data sources** (information from multiple sources is more powerful than information from any single source)

• **A knowledgeable, engaged user community** to help populate the database (crowd-sourced with a curated crowd.)

• **Ease of integration** with and use by other pieces of software - databases aren’t very useful if no one uses them. A UI is just another application.

• **Economic viability** – cheap is good, and persistence requires longevity, which requires money
What I Didn’t Want

- **Rights Information** - context dependent, varies by time and country, and some of it is secret and wrapped up in complicated legal documents.
- **People thinking they ‘own’ the metadata or the ID** -- metadata for identifying audiovisual works is freely available in lots of inconvenient ways, but it is available, so why should a record be ‘owned’?
- **Restrictions on read access to the metadata** - it’s only useful if people can use it as they wish.
- **Uncontrolled additions and modifications** -- when dealing with uniqueness, sloppy metadata is death.
- **Only commercially important works** - that would give a woefully incomplete picture of the world of film and TV, and preclude lots of interesting projects and applications.
- **Too much legalese** - the user agreement is only a couple of pages long, which is remarkable for something that is the offspring of Hollywood and Silicon Valley.
What is EIDR – spiffy marketing version

What EIDR is

- Global registry for unique identification of movie and TV content
- Designed for automated machine-to-machine communication
- Flexible data hierarchy down to the product & SKU level, incl. edits, clips, composites, encodings, and relationships

What EIDR is Not

- Profit-making
- Rich commercial metadata
- Ownership or rights information
- US-only

EIDR Purpose

- Make digital distribution competitive
- Help reduce costs
- Improve collaboration and automation across multiple application domains & platforms
- Enable new businesses and create new efficiencies

EIDR Technology Summary

- Interoperable, standards-based infrastructure
- Built on ISO Digital Object Identifier (DOI) standard
- Application integration through public APIs and schemas, freely available SDK for members
- Efficient infrastructure for new and existing applications
Example EIDR movie hierarchy with multiple versions
Example EIDR episodic hierarchy
What Metadata Is Included

• Enough descriptive metadata to distinguish one work from another
  • Different episodes of a Series, remakes, director’s cuts, dubbed languages, etc
  • Different requirements at different levels of the hierarchy
  • Factual, not interpretive = no genres or plot summaries, e.g.
    • Some exceptions (e.g. early Actualities)

• Links to other EIDR IDs
  • Containing Series, original abstract work, items included in a retail compilation, etc
  • We try pretty hard to make this concrete, not marketing or opinion (so we don’t do ‘franchises’, for example)

• Note on EIDR IDs
  • Standard form is 10.5240/C1B5-3BA1-8991-A571-8472-W
  • Lots of other formats defined for various applications
  • Resolution and content negotiation through the DOI proxy
What’s in an external identifier?

• The ID, in whatever its basic, native form happens to be
  • Currently only for other audiovisual works
  • Significant interest in covering other things (posters, scripts, reviews, miscellaneous ephemera)
  • …but none of those have good identifiers (yet)

• A type (IMDB, ISAN, DOI, etc)
  • If it’s not a ‘first-class type’, the type is ‘Proprietary’
  • Proprietary types have a domain within which they are valid, e.g. bfi.org.uk, bfi.org.uk/gifford, Disney.com, Disney.com/MPM, etc

• A resolvable URL for that ID, if we know how to make one
  • Generated on the fly from a published list of patterns
  • Some ID types have none, some have more than one

• A relationship (optional)
  • IsSameAs, IsDerivedFrom, ContainsPartOf, etc
# Top Alternate IDs in EIDR

<table>
<thead>
<tr>
<th>Alt. ID Type</th>
<th>2018 (000)</th>
<th>Chg.</th>
<th></th>
<th>Alt. ID Type</th>
<th>2018 (000)</th>
<th>Chg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TiVo ID</td>
<td>483</td>
<td>39</td>
<td></td>
<td>Flixster-Rotten Tomatoes</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>Webedia/AlloCiné</td>
<td>404</td>
<td>11</td>
<td></td>
<td>The Movie Database*</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>IMDb</td>
<td>314</td>
<td>69</td>
<td></td>
<td>NBCU</td>
<td>92</td>
<td>38</td>
</tr>
<tr>
<td>CITWF</td>
<td>290</td>
<td>30</td>
<td></td>
<td>ČSFD*</td>
<td>87</td>
<td>87</td>
</tr>
<tr>
<td>Sony</td>
<td>275</td>
<td>13</td>
<td></td>
<td>Turner Classic Movies (TCM)*</td>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>ISAN</td>
<td>209</td>
<td>20</td>
<td></td>
<td>Baseline</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>ITV</td>
<td>174</td>
<td>14</td>
<td></td>
<td>European AVO</td>
<td>57</td>
<td>2</td>
</tr>
<tr>
<td>Wikidata*</td>
<td>130</td>
<td>123</td>
<td></td>
<td>Google Play</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>Warner</td>
<td>112</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* New top ID this year
Quick examples

• Le voyage dans la lune [Voyage to the Moon] ( Méliès, 1902 )
  • 10.5240/C79D-4B8B-36AF-B214-03C0-H

• Frankenstein (Edison, 1910)
  • 10.5240/4375-1EB2-6CCD-DC73-B77F-V

• Game of Thrones
  • 10.5240/C1B5-3BA1-8991-A571-8472-W

• Blade Runner
  • https://doi.org/10.5240/EA73-79D7-1B2B-B378-3A73-M
First Application: Digital Supply Chain (simplified)

**Studio**
- Avail
- Metadata
- Fulfillment Metadata
- Media Files

**Retailer**
- Buy/Rent
- Entitlement Database
- Library Browse/Fulfill
Second application: Getting it right – double-shot movies

• Some movies aren’t dubbed
• Scenes are re-shot at more or less the same time in a different language sometimes with some of the actors different
• These aren’t edits of the same movie – they meet the definition of a separate work
  • Common in the 1930s
  • Still done for some Indian productions (Tamil/Hindi, for example)
Double-shot examples

- [https://doi.org/10.5240/BD8D-8F89-8F75-FE28-7010-M](https://doi.org/10.5240/BD8D-8F89-8F75-FE28-7010-M)
  - Murder! 1930 GB Double-shot in English (this version) and German.

- [https://doi.org/10.5240/C264-EC88-AFA1-2EC2-9B28-Z](https://doi.org/10.5240/C264-EC88-AFA1-2EC2-9B28-Z)
  - Mary 1931 GB DE Double-shot in English and German (this version).

- [https://doi.org/10.5240/1388-8D7E-42D2-7147-D5DB-L](https://doi.org/10.5240/1388-8D7E-42D2-7147-D5DB-L)
  - S.O.S Iceberg 1933 US DE Double-shot in German and English (this version).

- [https://doi.org/10.5240/9162-6940-4DC3-ABF9-A67B-0](https://doi.org/10.5240/9162-6940-4DC3-ABF9-A67B-0)
  - S.O.S. Eisberg 1933 DE US Double-shot in English and German (this version).

- [https://doi.org/10.5240/C97B-42DD-BF23-B3FF-4A8B-9](https://doi.org/10.5240/C97B-42DD-BF23-B3FF-4A8B-9)
  - Raavan 6/18/2010 IN Shot simultaneously with the Tamil-language version, Raavanan.

- [https://doi.org/10.5240/F635-4E44-475B-158B-9FF4-Z](https://doi.org/10.5240/F635-4E44-475B-158B-9FF4-Z)
  - Raavanan 6/18/2010 IN Shot simultaneously with the Hindi-language version, Raavan.

- [https://doi.org/10.5240/13D5-090F-CA5A-A590-CE47-5](https://doi.org/10.5240/13D5-090F-CA5A-A590-CE47-5)
  - Mumbai Express 4/15/2005 IN Double-shot in Hindi (this version) and Tamil.

- [https://doi.org/10.5240/2492-A8ED-46AE-7631-40F1-H](https://doi.org/10.5240/2492-A8ED-46AE-7631-40F1-H)
  - Mumbai Express 4/15/2005 IN Double-shot in Hindi and Tamil (this version).
Third Application: getting all of it -- This Modern Age

• The Rank Organisation’s answer (1946-1954) to The March of Time (by Time, Inc.)
  • Not the 1931 MGM film (that’s https://doi.org/10.5240/DE76-BA99-3701-6237-6BCE-I )
  • ITV has most of the Rank catalogue; BFI has some of it; CITWF has metadata for some of it

• All three had partial data
  • Combining data from all the sources in the EIDR records gives better overall information
  • All parties can update their records when they want to

• And they can also talk to each other about possible collaborations
  • Re-release with supplemental material from BFI, e.g.
  • Should also make later researchers’ lives easier
  • See https://doi.org/10.5240/E051-49A0-94DB-28CC-9F5F-Z for the results
Fourth Application: Linked Data

- Insert Tim Berners-Lee’s definition of linked open data here
- Problem is the world doesn’t work that way
- But the core of it is following /resolving identifiers to get to more stuff
- So MovieLabs built an ontology
  - Much more data that the basic stuff in EIDR
  - Treat provenance as an essential item
  - Heavy emphasis on which country or region the data applies to/came from
- Put EIDR data into the form the ontology wants, then....
  - Follow alternate IDs in EIDR to other sources, and adding that data
    - Wikidata (ontological, but difficult) and IMDB/S3 (not an ontology but very well structured and documented)
    - BFI (not an ontology, but an API, an XSD schema, and mappable controlled vocabulary)
  - Then several more sources
  - Build some applications
Extras

• Links to non audiovisual ‘stuff’
  • Books, plays, theme park rides, ....
• Editorial comments (reviews, ratings, rankings)
• Expanded person records
• Lots of locations (production, filming, setting)
• Artwork
• Reviews
• Consumption
Demo, if time and the fact that it’s a prototype permit....

- GraphQL queries
- UI
- A bit of machine learning or clever math (depending on your buzzword breaking point)
Why EIDR is Like Every Other PID

• Building a community of practice
  • Or ‘Why can’t our users get their metadata act together?’ and ‘They want to do WHAT?’
  • Connect communities that may not know what they have in common
  • Format wars

• Where does the identifier end and the application begin?

• Shared terms and vocabulary.
  • What does ‘version’ mean? What do ‘the same’ and ‘derived from’ mean?

• Identifiers for people and organizations

• What happens when two subcommunities disagree (vigorously) on the ‘best’ way to deal with a corner case?

• How do we talk to people about this stuff?
  • Explain why, not how
  • Deal with misconceptions and myths
Two messages from the past

...we are able to be the Secretaries, the interpreters and preservers of the memorials of our ancestors.


...I had the greatest intimacy with...the whole sett of learned men...and by having recourse to their libraries I arriv’d to a considerable degree of knowledge & equal reputation