Presentation Abstract

A horizon-scanning exercise, describing an ambitious vision for a scalable, dynamic film archive collections information architecture which would enable the community to:

- search across all peer collections, to interrogate holdings based on shared data models and API architecture, linked open data frameworks and shared unique identifiers
- surface a dynamically aggregated collections search and display offer to public users, based on the architecture outlined above
- automate comparison of holdings for specific films, directors, actors, production companies and genres across multiple archives, using Linked Open Data principles and unique identifiers

There are three major developments in the film archive domain, which make possible a radical transformation of collections information sharing and access:

1. Shared film metadata model: EN 15907, with the conceptual framework it embodies, makes it possible to understand our databases in like for like terms, for the first time
2. Collections Management Systems with open APIs: the growth of systems featuring API technologies enables new models of information exchange and querying
3. Unique identifiers for moving image works: the exponential growth of EIDR registrations, and their alternate ID cross-references, enables film collection systems to share information about the same film with confidence, and without labour

Together, these make possible a scalable aggregation of holdings metadata across all film archives, with no requirement for a central data store, or manual aggregation. Instead we can imagine applications which call APIs dynamically, to retrieve holdings using the unique ID as the key and the shared data model as the map.

Some warnings about this presentation...

Personal
- This is a vision, not the vision
- Archives may prefer to not expose their collections

Speculative
- This is not an explicit BFI ambition
- This is not a project plan

Theoretical
- This posits an ideal future, without reference to budgets or politics

Biased
- This is the vision of a Head of Data, not a Curator, Conservationist, Collections Manager or Lawyer

A vision for the future of film archive search

Make the machines:
  search all film archive collections and
  understand the results
  without requiring human work
  (type select export worry   explain)
Make this infrastructure:
  scalable
  affordable
  decentralised
  automatic

How is this different from **European Film Gateway**?

E.F.G.

- Centralised database / portal
- Actively / manually created by collection owners
- Iterative updates
- Selected / curated

This

- Decentralised
- Automatically created by systems
- Dynamic updates
- Self-selecting / un-curated

How is this different from European Film Gateway?

1. **Centralised versus decentralised:** EFG offers a centralised portal, with collection owners adding records to the central store, but this is a model for aggregating in a decentralised form, local systems joining up

2. **Active / manual versus passive / automatic:** EFG is created by collection owners consciously and actively uploading metadata and content, but this is works in the background, automatically created on the fly by systems

3. **Iterative updates versus dynamic updates:** EFG updates in iterations when contributors undertake their work, but this updates dynamically as archives create data in their systems

4. **Selected / curated versus self-selecting / un-curated:** EFG offers a selection of digital collections, but this is all collections, analogue and digital, not a selection. As a result it is a warts-and-all, no human intervention model...

So how does it work?

A vision for the future of film archive search
The building blocks

1. Shared data model: EN 15907

- Work Manifestation Item
- Comparing like with like
- Inherited properties

www.filmstandards.org

Film identification - Enhancing interoperability of metadata - Element sets and structures

EN 15907 defines a metadata set for the comprehensive description of cinematographic works including the various incarnations it can assume during its lifecycle

What is **EN 15907**?

Full Title: Film identification - Enhancing interoperability of metadata - Element sets and structures

EN 15907 defines a metadata set for the comprehensive description of cinematographic works including the various incarnations it can assume during its lifecycle.

EN 15907 is an event-based model, and one of its major benefits is enabling clear, accurate description of where your collection objects sit in the lifecycle of the work: cinema releases, censorship activities, restorations, reissues, etc.

For more information, see filmstandards.org.

How does EN 15907 enable archive search?

It describes a hierarchical model, Work - Manifestation - Item, with an optional Variant: each level contains a defined set of data, describing abstract properties, contextual properties, and concrete properties, with inheritance.

How does that help, when machines speak to machines about film collections?

It means they compare like with like, with clearly defined expectations and parameters. The search machines expect to find Director, Production Company and Copyright Holder in a Work record, and the dates they find there relate to production and copyright events. They expect to find release date or restoration date in a Manifestation; and instead of Director, they expect to find such agents as Distributor, Restoration Producer, and Censorship Authority. In the Item they expect to find acquisition and accession dates, print dates or file creation and checksum dates for digital Items, and agents including Acquisition Source, Conservationist, etc.

…

How does EN 15907 enable archive search?

...

It means they can rely on inherited properties, not duplicated properties across all object records.

This builds storage and search index efficiency when scaled across millions of records. Filmographic properties from the Work (e.g., Director, Synopsis, etc.) are inherited by the Manifestations and Items, contextual properties from the Manifestation (e.g., type of lifecycle event such as restoration, cinema release, etc.) are inherited by the Items.

This is efficient and clear, for machines and humans.

A vision for the future of film archive search
The building blocks

EN 15907: a shared data model for describing moving image works and collections

Work
abstract entity

- Titles
- Dates (copyright / production)
- Categories (fiction / non-fiction)
- Part - Whole conditions (serial / standalone / component part)
- Content: Synopsis, Genre, Subject
- Agents: Cast, Credits, Rights holders

Manifestation
realisation, release, exhibition or distribution entity

- Titles
- Dates: release, transmission, distribution, creation
- Type: pre-release, theatrical, non-theatrical, transmission, home-viewing, internet, restoration, not-for-release, etc
- Format general: 35mm film, Digital Cinema, Blu-ray, etc
- Agents: Creator, Broadcaster, Distributor, Publisher
- Rights context: platforms, territories, dates

Item
physical or digital object

- Titles
- Dates: creation, acquisition, accession, de-accession, loan, transport
- Acquisition: source, meathod, funding context, conditions of access
- Format specific: 16mm BW Pos, 35mm Lavender Separation, ProRes422 HQ, etc
- Condition report: pristine, not for projection, heavy scratches, etc
- Storage location: home location, current location
- Conservation recommendations: urgent transfer required, relocate to sub-zero, etc

Work
abstract entity

- Titles
- Dates (copyright / production)
- Categories (fiction / non-fiction)
- Part - Whole conditions (serial / standalone / component part)
- Content: Synopsis, Genre, Subject
- Agents: Cast, Credits, Rights holders

Variant
abstract entity
(optional)

- Titles
- Dates (copyright / production)
- Categories (fiction / non-fiction)
- Part - Whole conditions (serial / standalone / component part)
- Content: Synopsis, Genre, Subject
- Agents: Cast, Credits, Rights holders

Manifestation
realisation, release, exhibition or distribution entity

- Titles
- Dates: release, transmission, distribution, creation
- Type: pre-release, theatrical, non-theatrical, transmission, home-viewing, internet, restoration, not-for-release, etc
- Format general: 35mm film, Digital Cinema, Blu-ray, etc
- Agents: Creator, Broadcaster, Distributor, Publisher
- Rights context: platforms, territories, dates

Item
physical or digital object

- Titles
- Dates: creation, acquisition, accession, de-accession, loan, transport
- Acquisition: source, meathod, funding context, conditions of access
- Format specific: 16mm BW Pos, 35mm Lavender Separation, ProRes422 HQ, etc
- Condition report: pristine, not for projection, heavy scratches, etc
- Storage location: home location, current location
- Conservation recommendations: urgent transfer required, relocate to sub-zero, etc

A vision for the future of film archive search
The building blocks

The search machines can look in the same place for the same type of data: it is predictable and structured, therefore you can build rules and logic to automate search, retrieval and display. It is based on a concept (events in the lifecycle of a work), but it enables very granular, controlled logic for applications.

This shows the hierarchical relationships in 15907 expressed as separate but linked records: and this shows a three level model or a four level model, including the optional Variant.

The EN 15907 data model is probably best achieved as a set of discrete, linked records, each level enforcing its own fields and value lists, and allowed relationships. But the data model is really a conceptual model, and it can be comfortably achieved in a single level, flat database architecture.

A vision for the future of film archive search
The building blocks

EN 15907: a shared data model for describing moving image works and collections

Work

Manifestation

Item

properties expressed
in one record,
with abstract, contextual
and object data
stored on a single
hierarchy level

- Titles
- Dates: copyright, production, release, object creation, object acquisition / accession, de-accession, loan, transport
- Categories: fiction / non-fiction
- Part - Whole conditions: serial / standalone / component part
- Content: Synopsis, Genre, Subject
- Agents: Cast, Credits, Rights holders
- Dates: release, transmission, distribution, creation
- Type: pre-release, theatrical, non-theatrical, transmission, home-viewing, internet, restoration, not-for-release, etc
- Format general: 35mm film, Digital Cinema, Blu-ray, etc
- Agents: Creator, Broadcaster, Distributor, Publisher
- Rights context: platforms, territories, dates
- Acquisition: source, method, funding context, conditions of access
- Format specific: 16mm BW Pos, 35mm Lavender Separation, ProRes422 HQ, etc
- Condition report: pristine, not for projection, heavy scratches, etc
- Storage location: home location, current location
- Conservation recommendations: urgent transfer required, relocate to sub-zero, etc

A vision for the future of film archive search
The building blocks

The previous diagram shows how many moving image archives store their collection records: the abstract or filmographic units of information combined with the contextual and the object description units of information, in a flat database model.

One of the many drawbacks of this approach is the duplication of filmographic data across all object records (i.e., if you want to know the Director or the Synopsis, you have to store it in every record). This creates inefficiency in data storage and retrieval.

Another drawback is that it is more difficult to achieve a clear display of context: which objects of many are related to the same restoration, or the same cinema release, or the same censorship event. That may be captured in a text field, but it's not displayed structurally: the Items are not grouped together.

A vision for the future of film archive search
The building blocks

1. Shared data model: EN 15907	2. Systems with open APIs
Work Manifestation Item	Machines speaking to machines
Comparing like with like	Exposing datasets over internet
Inherited filmographic properties	Embedding data model EN 15907
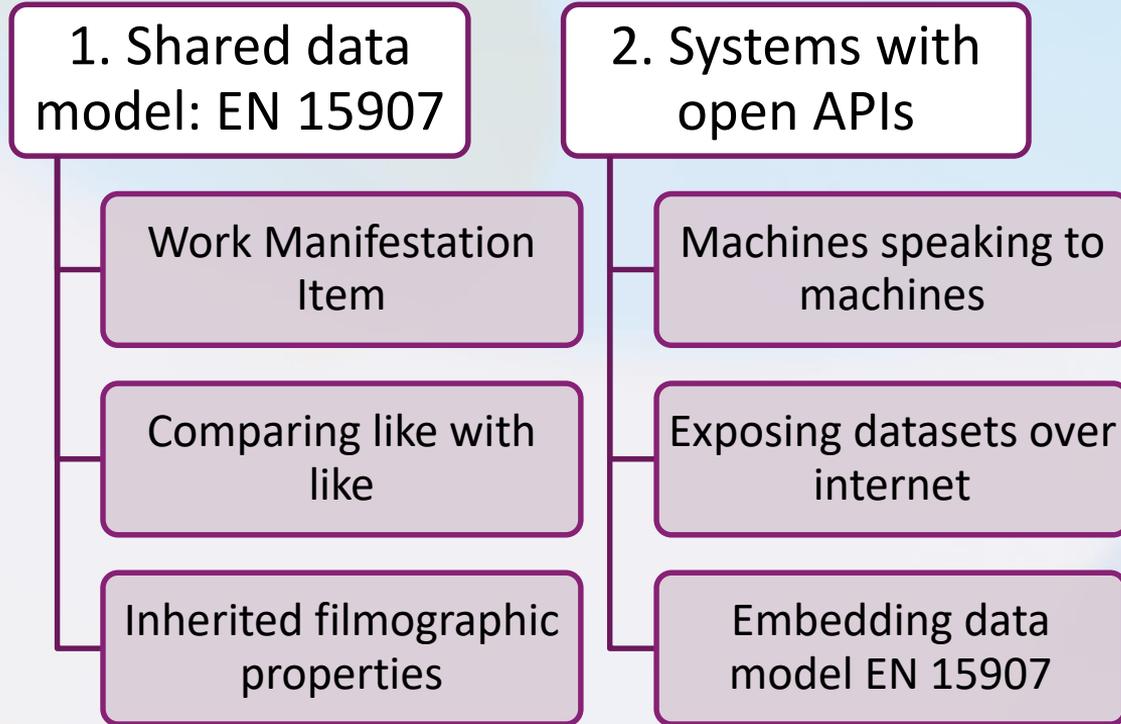
A vision for the future of film archive search
The building blocks

What is an API?

An API is a technology that lets systems interact with other systems, datasets, digital files, social media, networks, websites, fridge freezers, air conditioning systems, smart TVs, toothbrushes…

Per Wikipedia: In computer programming, an application programming interface (API) is a set of routines, protocols, and tools for building software applications… An API can assist otherwise distinct applications with sharing data, which can help to integrate and enhance the functionalities of the applications. APIs often come in the form of a library that includes specifications for routines, data structures, object classes, and variables. In other cases, notably SOAP and REST services, an API is simply a specification of remote calls exposed to the API consumers. Source: en.wikipedia.org/wiki/Application_programming_interface

A vision for the future of film archive search
The building blocks

A Generic API Example
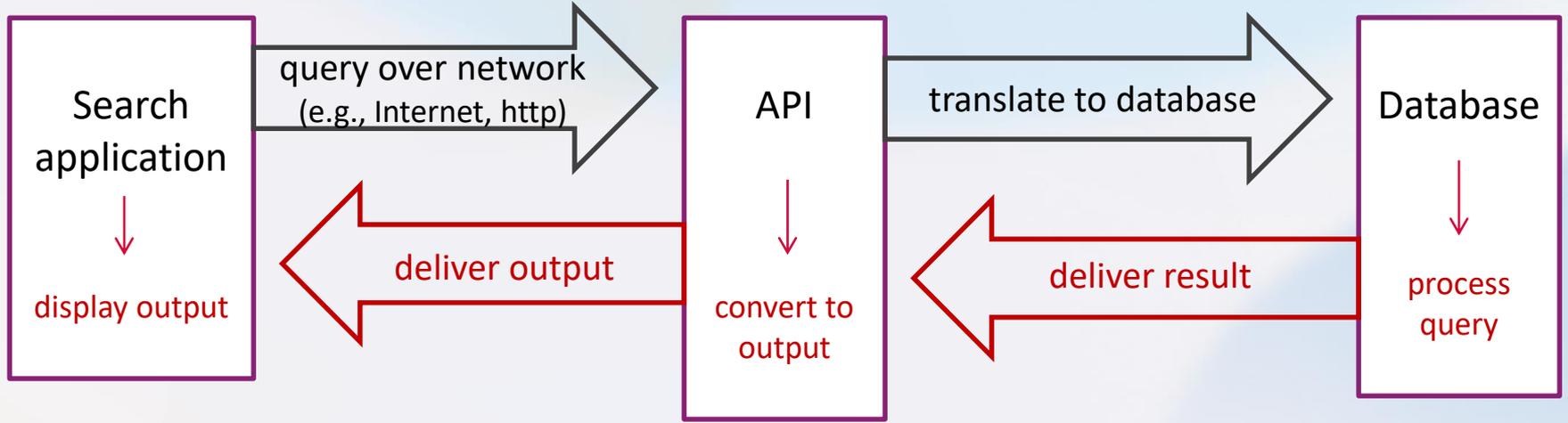
A vision for the future of film archive search
The building blocks

A Specific API Example

Adlib API response to the query:
database = works & search = title = 'red shoes' and credit name = 'powell, michael'

The BFI uses Adlib, supplied by Axiell. It has a very easy to use, documented API (api.adlibsoft.com), and you can ask it questions from a Web browser, API client, Excel document, Web site, etc.; and it will give you back the results as raw XML or JSON, or formatted as you like, using HTML, CSS, JavaScript, etc.

The following slide shows the Adlib API response to the above query when submitted to the BFI's Collections Information Database system.

A vision for the future of film archive search

The building blocks

The API's Response (in raw XML)

```xml
- <adlibXML>
  - <recordList>
    - <record priref="150033051" created="2011-08-15T12:02:30" modification="2014-08-28T15:38:01" selected="False">
      - <Content_subject>
          <content.subject>Ballet</content.subject>
        </Content_subject>
      - <Content_subject>
          <content.subject>Ballet schools</content.subject>
        </Content_subject>
      - <Description>
        - <description>
            ©. Independent Producers Ltd.; [logo] J. Arthur Rank presents; [logo] a production of the Archers; released through General Film Distributors; [end] a production of the Archers
          </description>
        </Description>
      - <Description>
        - <description>
            Fantasy of a young girl who achieves fame in a ballet company dancing in the story of Andersen's 'Red Shoes'. Torn between love for a composer and artistic devotion to the ballet master, she dances to her death..
          </description>
        </Description>
          <priref>150033051</priref>
      - <Title>
          <title>Red Shoes</title>
        </Title>
      - <Title_date>
          <title_date_start>1948-07-22</title_date_start>
        </Title_date>
      </record>
    - <record priref="150274683" created="2011-08-15T21:54:27" modification="2015-03-30T17:18:53" selected="False">
      - <Content_subject>
          <content.subject>Music</content.subject>
        </Content_subject>
      - <Description>
        - <description>
            Work consists of the ballet scenes taken from 'The Red Shoes' (1948).
          </description>
        </Description>
          <priref>150274683</priref>
      - <Title>
          <title>Red Shoes</title>
        </Title>
      - <Title_date>
          <title_date_start>1948</title_date_start>
        </Title_date>
      </record>
    </recordList>
  - <diagnostic>
```

A vision for the future of film archive search
The building blocks

The API's Response (reformatted)

UFH grouping	Selected by	CID object number	CID priref	Title	Title Language	Date	Work Type	Category	Country	Production Company	Director	©	Cast / Participants	Genre	Subject	Synopsis	Production notes	Colour / BW	Duration	Sound	Language	Language usage
-all-	-all-				-all-	-all-	-a	-all-	-all-					-al				-al		-a	-all-	-all-
unspecified	BFI	42632	150033051	Red Shoes, The (Original)	English	1948-07-22 (Release)	Film	Fiction	United Kingdom	Independent Producers / Archers Film Productions / Rank Organisation, The / General Film Distributors	Michael Powell / Emeric Pressburger	empty	Anton Walbrook: *Boris Lermontov* / Marius Goring: *Julian Craster* / Moira Shearer: *Victoria Page* / Robert Helpmann: *Ivan Boleslawsky* / Léonide Massine: *Grischa Ljubov* / Albert Bassermann: *Ratov* / Ludmilla Tcherina: *Irina Boronskaja* / Esmond Knight: *Livingstone 'Livy' Montague* / Jean Short: *Terry* / Gordon Littman: *Ike ...*	Drama	Ballet / Ballet schools	©. Independent Producers Ltd.; [logo] J. Arthur Rank presents; [logo] a production of the Archers;...Fantasy of a young girl who achieves fame in a ballet company dancing in the story of Andersen's 'R...	empty	Colour	136	Sound	English	Dialogue (original)
unspecified	BFI	268592	150274683	Red Shoes, The (Original)	empty	1948 (Production)	Film	Non Fiction	United Kingdom	C.B.A.	Michael Powell / Emeric Pressburger	empty	empty	empty	Music	Work consists of the ballet scenes taken from 'The Red Shoes' (1948)....	This is the film consisting of the ballet scenes.	empty	20	empty	empty	empty

A vision for the future of film archive search
The building blocks

The API's Response (on a Web page)

*Note that there are two Red Shoes by Michael Powell in 1948.

A vision for the future of film archive search
The building blocks

An API – like the Adlib API, which is in use by more than 6 film archives in Europe and USA – plays these key roles:

1. It exposes film archive collections data to querying over the internet (HTTP), without risking any database violation (it is a translator and firewall between the outside world and the database back end)

2. It exposes those datasets to a structured and documented set of search operations (command library) which can embed EN 15907's principles without having to change the database structure (the API is a translator)

3. It delivers back to the search application machine processable information in XML or JSON, and it can transform that using technologies like XSLT to any form that the display requires

The API doesn't care what the search application is: it can be a website, an iPad app, an Excel spreadsheet, Access database, Word document, etc. As long as the application has Internet access it can ask the API a question, receive the response, and do what it likes with the response: display it in a Web browser, store it in a file, email it, tweet it, or post it to Facebook.

A vision for the future of film archive search
The building blocks

1. Shared data model: EN 15907	2. Systems with open APIs	3. Unique IDs: EIDR
Work Manifestation Item	Machines speaking to machines	
Comparing like with like	Exposing datasets over internet	
Inherited filmographic properties	Embedding data model EN 15907	
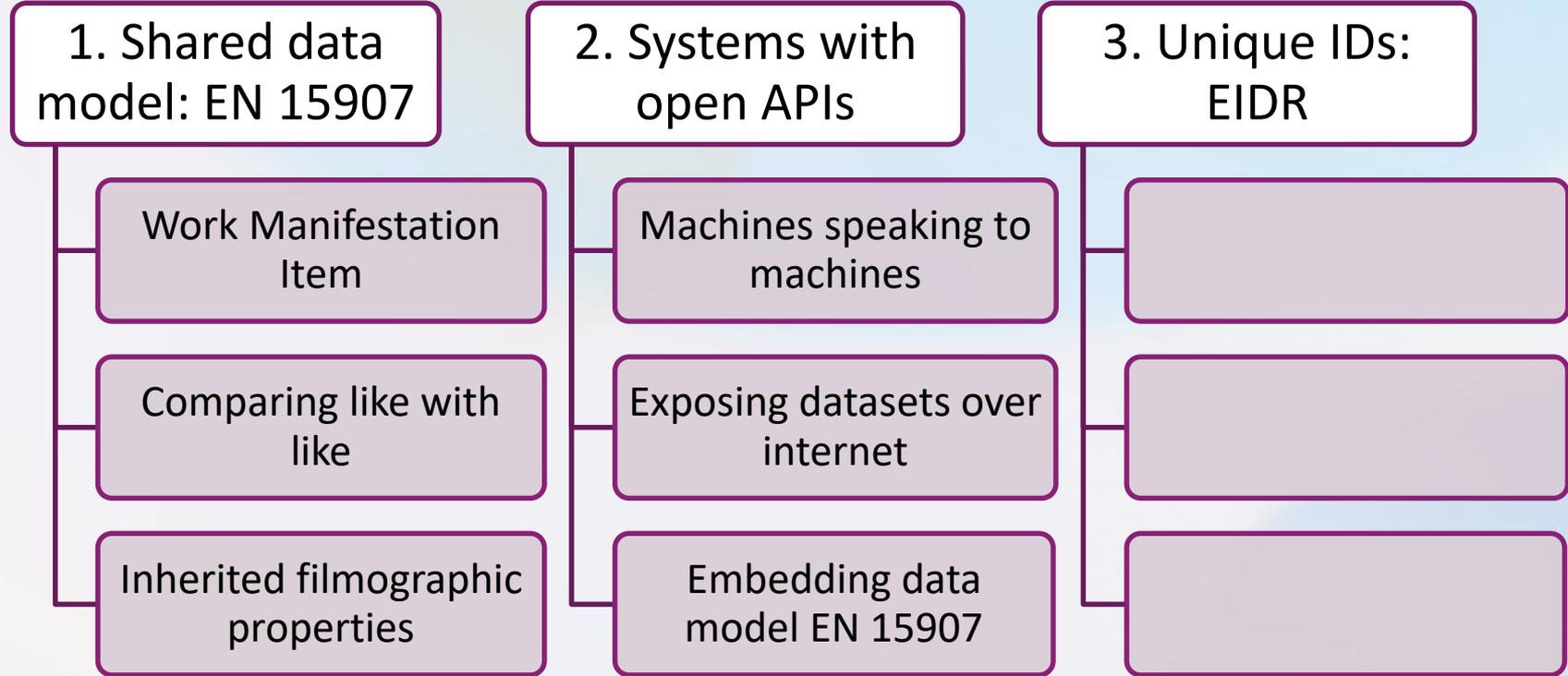
A vision for the future of film archive search
The building blocks

A key identifier in this domain is EIDR, a global ISO 26324:2012 standard ID Registry for the media and entertainment industry.

EIDR IDs are to audiovisual assets (movie, TV, radio, and Web) as ISBNs are to books and UPCs/EANs are to consumer products.

The use of globally-unique, curated audiovisual asset identifiers is reaching critical mass now, with all the major film, TV and Internet companies using them to achieve economies of scale in digital moving image ecosystems such as video-on-demand, e-commerce, etc.

ISO: International Organization for Standardization – www.iso.org

ISO 26324:2012: Digital Object Identifier (DOI) – www.doi.org

EIDR: Entertainment Identifier Registry Association – eidr.org

A vision for the future of film archive search
The building blocks

How do globally-unique identifiers (e.g., EIDR) promote this vision for the future of film archive search?

Remember the Red Shoes? (Our API search returned two results when we searched for Red Shoes by Michael Powell in 1948.)

Even when you know director, title, and date, the results of film searches can be unexpected: two works titled *Red Shoes*, with Powell as director, dated 1948 (one being the full film, the other being the ballet scenes only).

Titles, dates, and names are not enough to let machines properly speak to each other about film works – although they are a good place to start. For the machines to speak quickly to each other, with confidence about the answers, they need globally unique identifiers.

A vision for the future of film archive search
The building blocks

De-duplication and uniqueness

A vision for the future of film archive search
The building blocks

De-duplication and uniqueness

The previous example from the BFI's Adlib system shows the Film Work record where we have a field for an EIDR identifier (10.5240/7A8F-736F-2022-236F-C77E-Q).

The Red Shoes' EIDR ID is globally unique and machine processable. This lets anyone with the EIDR ID compare their database with ours, establish what we know about *The Red Shoes* that they don't, and, more importantly, what we hold in our collection.

*The EIDR record for *The Red Shoes* also contains more than 35 other identifiers, allowing ID cross-reference and direct connections to other organizations including Amazon, Česko-Slovenská filmová databáze, Cinémathèque Québécoise, ISAN, ITV, Netflix, Swedish Film Institute, Sony Pictures, and Warner Bros., many of which provide work-specific Web pages hyperlinked from the EIDR Registry.

A vision for the future of film archive search
The building blocks

1. Shared data model: EN 15907	2. Systems with open APIs	3. Unique IDs: EIDR
Work Manifestation Item	Machines speaking to machines	Unambiguous identification of works
Comparing like with like	Exposing datasets over internet	Machine processable identifiers
Inherited filmographic properties	Embedding data model EN 15907	Automation of comparison
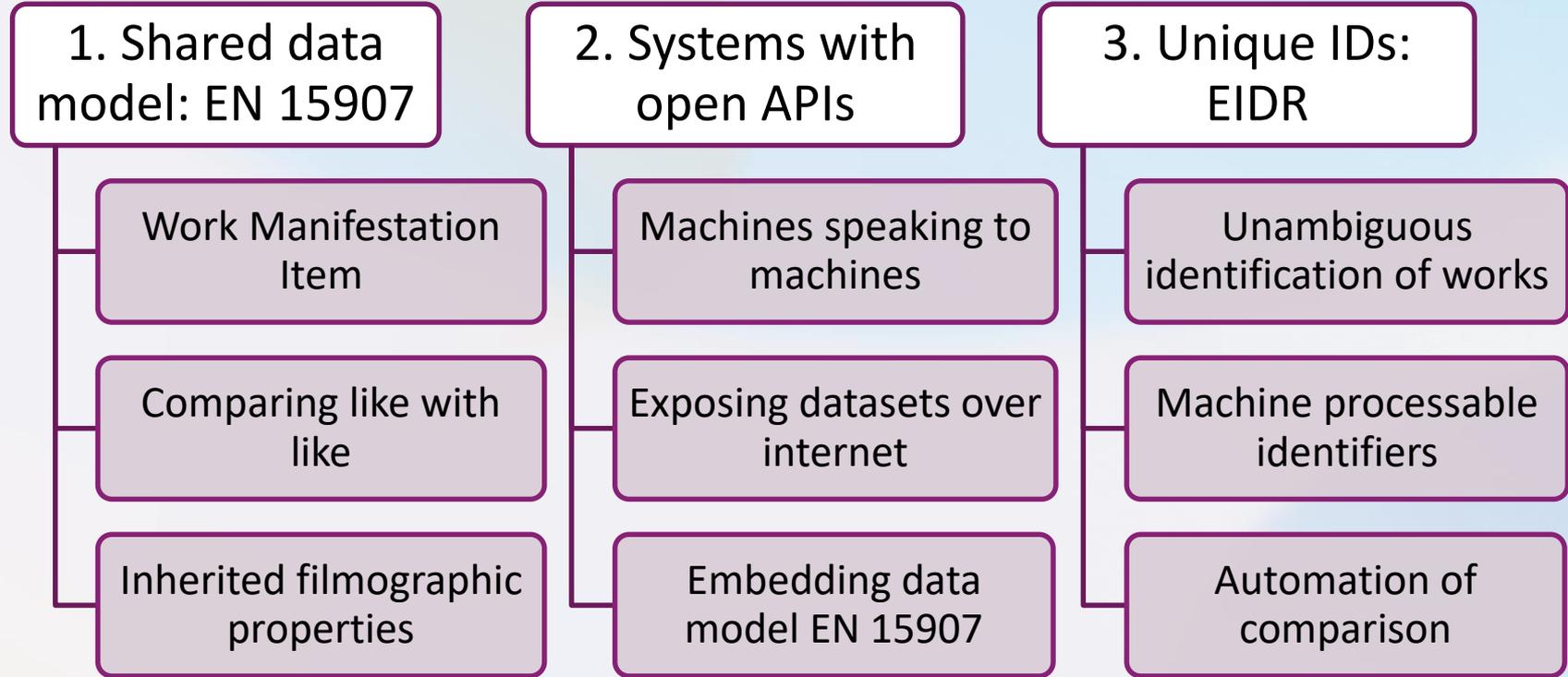
A vision for the future of film archive search
The building blocks

Globally-unique, curated identifiers offer:

1. **Unambiguous identification of works:** no room for confusion because of matching titles, dates, names. The deduplication and uniqueness has been achieved in the EIDR Registry.

2. **Machine processable identifiers:** machines can use the globally unique identifier to establish very quickly and without any debate, that this thing in domain 1 (e.g., the BFI) is the same as that thing in domains 2 to 50 (e.g., all other film archive databases).

3. **Automation of comparison:** the unique identifier allows applications to query APIs and establish very quickly and easily what is held by the organisation for this work. EN 15907 is the conceptual model for that comparison, but the unique ID is the fuel which makes it run quickly and without anomalous results

So, how might it look, in practical application?

Imagine the search for the Red Shoes if it was able to interrogate every film archive database API in the world, on the fly, silently, in seconds, using the unique identifier; understand the results based on EN 15907, and display that to the film archivist, the researcher, the rightsholder, to Martin Scorsese...

A vision for the future of film archive search
The building blocks

Deduplication and uniqueness

A vision for the future of film archive search

The BFI Collections Search result for The Red Shoes. But in addition to symbols indicating the types of content held by the BFI for the Red Shoes, imagine you could see which FIAF member archives in the world held related materials.

A vision for the future of film archive search

Imagine you could then filter by pre-print material, 35mm prints, DCP, DCDM, DPX; by original theatrical release, 1980 reissue; by condition of the elements, by access cost.

Imagine you could navigate directly to an email to the responsible curator or collections access department. Or imagine you could view a proxy online, directly, as archives digitise their collections and store in MAM systems.

Imagine you could undertake the same search not just by work, but by director, art director, scriptwriter; or by subject, genre, country, decade of production, date of printing.

Imagine the results you obtain are organised by work, manifestation, item, guaranteed accurate by unique identifiers, displayed in seconds.

A vision for the future of film archive search

Make the machines:
  search all film archive collections and
  understand the results
  without requiring human work
  (type select export worry   explain)
Make this infrastructure:
  scalable
  affordable
  decentralised
  automatic

A vision for the future of film archive search

Let the human beings:
  find all versions of all films in all collections
  understand the results
  proceed with projects

Let the organisations:
  drive down access costs, create revenue
  justify and defend their collecting activities
  improve their collections
  keep collecting!

A vision for the future of film archive search

The outcome would be to unlock the world's moving image heritage, making our collections transparent to our peers at the click of a mouse, driving down the real terms costs of access and reuse of the collections.

I'll show you mine, if you show me yours: A vision for the future of film archive search

Stephen McConnachie,
Head of Data
Collections & Information, BFI

stephen.mcconnachie@bfi.org.uk
@mcnatch