The Power and Promise of Identification

The Benefits of Universal Identification and the 21st Century Audiovisual Archive

Authors
Richard W. Kroon, Director of Engineering, EIDR Association
Raymond Drewry, Principal Scientist, MovieLabs
Andrea Leigh, Moving Image Processing Unit Head, Library of Congress
Stephen McConnachie, Head of Data and Digital Preservation, BFI

"Archives constitute the memory of nations and societies, shape their identity, and are a cornerstone of the information society."

— International Council on Archives

"Let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident."

— Thomas Jefferson

"I rather think that archives exist to keep things safe, but not secret."

— Kevin Young

Synopsis
This paper lays out many of the advantages to be gained by audiovisual libraries and archives (memory institutions) when they join together to create service provider networks and how the application of universal identification based on the Digital Object Identifier (DOI) system can make this process easier, cheaper, and more effective while also opening up other opportunities for automation including universal search; academic citation and cross-reference; and the publication of Linked Open Data (LOD).

1 International Council on Archives. "Mission, Aim and Objectives."
2 Thomas Jefferson to Ebenezer Hazard. February 18, 1791.
3 Young, Kevin. “Ghost Story from a Haunted Conscience.”
**Introduction**

Audiovisual archives, whether part of a larger library or existing as an independent entity, are memory institutions that preserve and publish a large and growing portion of our collective cultural heritage, represented by a wide variety of audio, visual, and collateral assets recorded on a bewildering mix of disk, tape, film, paper, and other media. A physical card catalog can no longer support a memory institution’s tripartite mission of collection, preservation, and access.

The primary value of a memory institution is a function of its collections (physical and digital assets), catalogs (indexes of holdings and related assets), and search/retrieval services. All institutions, whether public or private, physical or digital, commercial or non-profit, must make their collections available to the communities they serve. To do this, they must know what they have, where it is, and communicate this to those who need it. Ferdinand Columbus, Christopher Columbus’ son and one of history’s great librarians, referred to any collection that is not properly cataloged and indexed as “dead.”

Every action carries a transactional cost, born in proportion by each institution and its users. Even non-profit institutions have a financial bottom line. To remain viable, institutions must reduce their transactional costs while expanding their audience to increase revenue or amortize costs across a wider base. This means an increasing reliance on automation and machine intelligence: automation is not possible; analytics are not meaningful; machine learning cannot return useful results; and big data quickly become overwhelming without effective and economical universal identification.

There are several types of identification, many of which coexist in libraries and archives today: shelf numbers/directory paths; classification and cataloging systems; inventory indexing; statistically unique identifiers; and curated, globally-unique identifiers. Traditional cataloging continues to play a central role, but cataloging relies on human expertise, leading to variation that limits automation, multi-institution workflows, and global cross-reference. By adopting a system of universal identification, memory institutions stand to gain considerable advantage, not the least of which is a practical mechanism for universal search.

**A Stroll through the Library Stacks**

As anyone who has ambled through a library will know, order is everything. The ways in which books can be organized multiplies...

---

rapidly as the collection grows, and each shows the universe in a slightly different light.\(^5\)

The works held by audiovisual memory institutions go far beyond the shelves of dusty books or climate-controlled vaults filled with film cans that people may imagine when they think of a library or an audiovisual archive. Even within a particular institution, a single cataloging system may not be sufficient to capture the wide variety of assets that must be identified, described, and categorized. For example, the British Film Institute uses an ISAD(G)-compliant system\(^6\) to catalog documents\(^7\) and Spectrum (a UK museum collection management standard)\(^8\) for image-based assets.\(^9\)

Many libraries rely on MARC\(^{10}\) records to catalog their collections in a standardized and interchangeable electronic format, but MARC records do not guarantee unique identification. For example, the GND project of the Deutsche Nationalbibliothek and German Library Networks seeks to redress this MARC limitation and “consign to the past problems arising from different formats, parallel storage of data records and different rules for descriptive and subject cataloging.”\(^{11}\) Thankfully, identification is a more constrained effort than traditional cataloging, and therefore achievable across an institution’s collections and across multiple institutions.

One of the best known identification systems is binomial nomenclature, “probably the most persistent identifier system to date.”\(^{12}\) This is the formal system for naming species using a two-part name (\textit{homo sapiens}, \textit{camellia sinensis}, \textit{canis familiaris}, and so on.) Carl Linnaeus formalized the system in \textit{Species Plantarum} (1753).\(^{13}\) The system has evolved since then, and has gone through various phases of development, but it is now the recognized standard for identifying living or once-living things. As an example, consider \textit{Galium Odoratum} – sweet woodruff in English and Waldmeister in German. A botanical garden, plant nursery, or biodiversity survey only needs to say “Galium Odoratum” and there will be absolutely no ambiguity about what is

---


\(^{6}\) International Council on Archives. “ISAD(G): General International Standard Archival Description.”

\(^{7}\) In their broadest sense.

\(^{8}\) Collections Trust. “Spectrum.”

\(^{9}\) All assets are cross-linked to the relevant moving image work(s) and to person(s), including author, artist, subject, donor, rights holder, etc.

\(^{10}\) Bibliographic: Machine-Readable Cataloging originally developed by the Library of Congress. Now MARC 21 (see Furrie, Betty. “Understanding MARC.”).

\(^{11}\) Deutsche Nationalbibliothek and German Library Networks, “Gemeinsame Normdatei (GND).”

\(^{12}\) Thompson, Henry S. “Understanding URI Ecosystems.”

\(^{13}\) Linnaeus, Carl. \textit{Species Plantarum}.
meant. This has taken years (centuries) of effort across multiple disciplines – botany, zoology, paleontology, bacteriology – but the results are there for anyone to use.

“Galium Odoratum” is an identifier. The metadata used to match against this identifier includes things like leaf shape, habitat, and flowers. A catalog entry contains much more information (how it is used, other names, and so on.) In some ways, the catalog preceded the identifier. The Herball, or Generall Historie of Plantes by John Gerard is one of the earliest English botanical compendia. Its entry on sweet woodruff starts:14

Woodrooffe hath many square stalks full of joynts, and at every knot or joynt seven or eight long narrow leaves, set round like a star, or the rowell of a spurre: the floures grow at the top of the stems, of a white colour and of a very sweet smell, as is the rest of the herbe, which being made up into garlands or bundles, and hanging up in houses in the heat of Summer, doth very well attemper the aire, coole and make fresh the place, to the delight and comfort of such as are therein.

It continues with a description of subspecies, plants that can be confused with it/are similar to it, habitat, time of flowering, names in other languages, attributes in the system of the four humours,15 and medicinal uses.16

For a more modern catalog entry, consider the one at the Royal Botanic Gardens, Kew:17

Figure 1. The Royal Botanic Gardens, Kew entry for Galium odoratum.

14 Gerard, John. The Herball, or Generall Historie of Plantes 2nd edition. (1636)
15 “in a meane between heate and drinesse”
16 “to be put into wine, to make a man merry, and to be good for heart and liver”: It is still used in Maiwein (May wine) in parts of Germany (see German Culture. “German Maiwein.”).
This is less stately than Gerard (times change), but still has a lot of information. Across the four tabs on the Web page it gives other binomial names by which it used to be known (alternate identifiers, as it were), some of which preserve Asperula, Gerard’s Latin name for the plant; known habitats; a brief description; links to other references about the plant; and, under “Further Information,” links to catalog information for Kew’s preserved specimens, including their Kew catalog identifiers. We have gone from the abstract (what is this plant?) to information about the specimens at this institution.

Commercial on-line plant catalogs will reveal even more, including historical background, advice on growing it, and recipes, but they are all keyed off the universal unique identifier Galium Odoratum. There are also numerous plant identification Web sites that take in metadata (size, shrubbiness, flower color, leaf arrangement, or even a picture) and identify the plant for you, giving the unique Linnean identifier and some catalog information.

In short, botanists, park rangers, climatologists, landscape architects, plant nurseries, and gardeners all know they are talking about the same thing – they understand the same unique, universal identifiers. The universal identifiers do not take the place of cataloging. Instead, they extend cataloging efforts, allowing them to interconnect, reference and build upon one another, and avoid ambiguity and mistaken identification.

Biodiversity is difficult to top in terms of variety when it comes to the challenges of accurate and useful identification, but the shelves, vaults, and cabinets of an audiovisual memory institution may still contain a bewildering diversity of work types to challenge the intrepid identifier. During a virtual stroll through an audiovisual collection, we’re quite likely to encounter many different asset types, each with a varying level of accompanying descriptive documentation that may depend on the asset type, data of acquisition, source, popularity, and perceived value. Each has its own unique identification challenges and potential solutions. This is all made more complicated by the accelerating trend towards digital collections, both born-digital and those converted from analog forms.

Many of these asset types can be identified today using an existing DOI Registry. Others, such as advertising ephemera (posters, lobby cards, publicity stills, etc.) and people, are under active development within the Media & Entertainment industry. Those that remain could easily be added to the DOI family of identifiers should an institution or consortium care to take

18 See “Appendix: Common Audiovisual Memory Institution Asset Classes.”
the initiative. The results would be universal identification for all asset types held by audiovisual memory institutions.

**In Which We Make the Case for Universal Identification**

Emperor penguins understand the importance of universal identification. Unlike other penguins, emperor penguins do not make nests, so mates and chicks cannot find each other based on nest location (penguin context metadata), and they all look nearly alike, so they can’t find each other based on appearance (penguin descriptive metadata). Instead, emperor penguins have a specialized vocal organ that allows them to voice two different sounds at the same time, so each penguin has a unique two-tone call that it uses to identify itself – a globally unique penguin identifier, or GUPID, if you will, used for both identification and search under conditions more challenging than at all but the most fraught institutions.¹⁹, ²⁰

Audiovisual libraries and archives come in different shapes and sizes from large public lending libraries to small commercial archives. They each have unique collections, missions, resources, and stakeholders but at their heart they are all memory institutions charged with the collection, preservation, and dissemination of our common cultural heritage. Their individual value can be dramatically increased by promoting their interconnectivity as part of a global network of memory institutions, which will in turn help ensure their viability and continued existence into the future.

A presentation to the International Federation of Film Archives (FIAF) called for a system of universal search, spanning archives around the world, which would enable the network of institutions to:²¹

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

Universal search is the ultimate benefit of establishing an interconnected network of memory institutions, but universal search first requires universal identification. For its part, universal identification enables many ancillary applications.

---

²⁰ Sturdy, Christopher B. et al. "Bird communication: Two voices are better than one."
²¹ McConnachie, Stephen. “I’ll show you mine, if you show me yours”
²² API: Application Programming Interface – a packaged set of tools for building software.
benefits including aggregating data from multiple sources, curating digital catalogs distributed across physical locations, and automating routine tasks. All of this leads to cost savings and increased engagement opportunities.

**Memory Institutions and the Network Effect**

No matter how valuable a memory institution is to its myriad stakeholders, it is of greater value when it is connected to other institutions. The more connections, and the easier they are made, the greater the increase in value for the participating institutions and the greater benefit to their respective stakeholders. Because, when the whole is greater than the sum of its parts, the parts have increased in value by association.

This is the network effect, first identified by AT&T in 1908 and later quantified c. 1980 by Robert Metcalfe.\(^{23,24}\) It can be defined as occurring when “the utility that a user derives from consumption of the good increases with the number of other agents consuming the good.”\(^{25}\) According to Metcalfe’s Law, the cost of connecting to a network is quickly dwarfed by the value that the participants gain, with each new participant bringing increased value to all the others.

![Network Value Gain (Metcalfe’s Law) vs. Institution Gain (by joining the network).](image)

Memory institutions are not simply connected nodes on a network. They derive much of their value from the people who use them and the communities that they serve, acting more like a social network than a computer network. So, while a networked institution’s value may increase

---

\(^{23}\) Romero, Jessie. “Network Effects.”


\(^{25}\) Katz, Michael and Carl Shapiro. “Network Externalities, Competition, and Compatibility.”

\(^{26}\) The value of a network is proportional to the square of the connected nodes, or \(n^2\).

\(^{27}\) On its own, an institution’s value is 1. By joining an interconnected network, the overall value of the network grows and each participating institution’s value grows by association.
according to Metcalfe’s Law based on the number of participating institutions, the potential value to its users increases according to Reed’s Law based on the total number of connected users. Consequently, the significant potential benefit an institution may enjoy by participating in an interconnected network is a fraction of the potential benefit to its users.

![Graph showing User value gain (Reed’s Law) vs. institutional value gain (Metcalfe’s Law).](image)

**Figure 3. User value gain (Reed’s Law) vs. institutional value gain (Metcalfe’s Law).**

Connected memory institutions and their users also benefit from a Marketplace Network Effect where more buyers (users) attract more sellers (institutions) and more sellers attract more buyers. In this arrangement, users benefit from greater choice (more content available in more forms from more providers) and institutions benefit from access to more potential users, which can lead to more transactions against a broader range of assets.

### Gaming the Pareto Principle

The Pareto Principle is named for Vilfredo Pareto, a 19th century Italian economist who first observed that 80% of Italy’s land was owned by 20% of its population. This later became known the 80/20 rule and can be generally applied as 80% of value comes from 20% of supply. So, on average, 80% of an institution’s value will come from 20% of its resources and 80% of its transactions come from 20% of its users.

Each asset has an intrinsic value, which may have no relation to its value as a cultural artifact. Broadcast pioneer David Sarnoff observed that the value

---

28 In 2001, MIT’s David Reed proposed that the value of a social network, where users can interact individually or as part of multiple, self-forming communities of interest, increases exponentially with the number of connected users, or $2^m$.


30 Ibid.

31 Kruse, Kevin. “The 80/20 Rule And How It Can Change Your Life.”
of a broadcast network is proportional to the number of viewers. Applied to memory institutions, Sarnoff’s Law would hold that the value of an asset is proportional to the number of users who access that asset.

Because of the social network effect (via Reed’s Law), an asset may have relatively little value in the aggregate but could still have very high value to a particular community of interest. The key to maximizing an institution’s cumulative value for all of its assets is reaching all of the communities of interest that might be served by any of an institution’s collections or catalogs. Each community of interest applying its own 80/20 rule across a connected network increases the likelihood that a particular asset held by a particular institution will be valued. And, since the various network effects act as value multipliers, the cumulative benefit is both broader and deeper than a simple 80/20 calculation might imply.

**Identifiers and the Identified**

Identifiers are at the heart of how data can be effectively published, retrieved, reused and linked. ... [They] are simply labels used to refer to an object being discussed or exchanged [the “identified”], such as products, companies or people. ... The lack of identifiers, or the poor use of them, stifles the power of information gained from linking multiple datasets together. Some of these shortcomings might be overcome using intelligent search and fuzzy matching, but the lower precision of these techniques means that the data never reaches its full potential and there is little incentive to drive improvement of precision over time.\(^\text{32}\)

Work titles are the most common form of identifier for audiovisual works found in an archive or library. They are not unique – more than one work may have the same title and the same work may be known by more than one title. Turning titles into unique identifiers requires title matching. Titles are often augmented with additional data points such as work type, creation or release date, cast or creators, synopsis or description, etc. to help ensure uniqueness.

This process is not limited to audiovisual works. Related assets, such as advertising ephemera, production artifacts, academic papers, and books each have their own title-matching issues and accompanying descriptive metadata provided in an attempt to mitigate these issues. Books, plays, and screenplays often use title and author in combination (the most common compound identifier used by libraries). Manuscript collections are identified

\(^{32}\) Beaumont, Stewart and Sir Nigel Shadbolt. “Creating Value with Identifiers in an Open Data World.”
by creator and descriptive title; popular music is commonly identified by the performer or composer and title of the album in combination; and many works don’t have settled titles. Actualities are identified by description. Posters often require promoted work title, size, print media, territory, and release. Production stills may have a production number, which requires the production company to resolve to a production title.

Looking at *Gone with the Wind* (1939) as an example, a memory institution’s catalog could include:

- Gone with the wind (Motion picture : 1939) [original theatrical release]
- Highsmith, Carol M., 1946- Likeness of Vivien Leigh in character as Scarlett O’Hara in the movie “Gone with the Wind,” at Madame Tussaud’s Museum in the Hollywood section of Los Angeles, California [graphic]
- Howard, Sidney Coe, 1891-1939. Gone with the wind [screenplay]
- Mitchell, Margaret, 1900-1949. Gone with the wind [novel]
- Steiner, Max, 1888-1971. Gone with the wind [sound recording]
- Steiner, Max, 1888-1971. Max Steiner papers
- Tretick, Stanley. Gone with the wind (Margaret Mitchell) [graphic]

These compound “work identifiers” are substantially more complex to manage than simple “work titles,” and there is nothing simple about managing work titles.

Like work titles and compound work identifiers, the descriptive metadata that further describe and identify a work can vary significantly by source and asset type and often over time. They require fuzzy logic to achieve a positive match, which often confounds automated systems. People are much better at fuzzy logic matching than machines, but they are also relatively expensive, slow, and inconsistent – particularly when dealing in large volumes, such as with an entire library’s catalog or the collections in an archive. Metadata matching can be automated to a point, but manual review cannot be eliminated as long as workflows rely on metadata to link and identify assets. The ultimate goal, then, should be to match once in an asset’s life and then link all future workflows with a shared identifier.

The computer scientist Andrew S. Tanenbaum once observed, “The nice thing about standards is that you have so many to choose from.”

---

33 Tanenbaum, Andrew S. *Computer Networks*. 2nd ed.
identifier do you choose? Which one is the identifier that will serve all purposes for all people?

The answer is: None of them.

No one identifier can serve every purpose and solve every need. Instead, we must look for an identifier system that best serves each particular purpose. For universal search and other multi-party applications, we need a service where machines can “search all … collections and understand the results without requiring human work” using an infrastructure that is scalable, affordable, decentralized, and automatic that will help organizations “drive down access costs, create revenue, justify and defend their collecting activities, [and] improve their collections.” To do this, we must start with a shared identifier that satisfies the Tolkien Rule:

One ID to rule them all,
One ID to find them,
One ID to bring them all and in the global ecosystem bind them.

So, we do not need an ID that replaces all others, but rather one that links all others. One that can be used across organizations and in multiple applications, connecting as needed to and from other identifiers coined for other purposes. One that can capture the relationships among various types of resources that use different content standards and metadata structures.

A useful identifier that satisfies the Tolkien Rule must come from an identifier system that can guarantee that the IDs will be:

- **Globally Unique** – a particular identifier references a particular thing. That thing could be a single asset or a collection of assets; it could be a physical thing, a digital representation, or an abstract reference; it could be something held in a collection or referenced in a catalog. Whatever it is and wherever it is, each identifier can only identify one thing.

- **Permanent** – once assigned, the identifier never goes away. It is always available for use in reference to the identified object and the association between the identifier and the identified never changes.

- **Consistent** – everyone always gets the same result every time. Cataloging systems are dependent on the cataloger; directory paths are dependent on the computer file system; statistically unique systems may deliver a different result each time they are applied. Inconsistency is the enemy of efficiency.

---

34 McConnachie, Stephen. “I’ll show you…”
35 After J.R.R. Tolkien from The Fellowship of the Ring.
36 Kroon, Richard W. “The secret to automating multi-party asset workflows.”
- **Open (Not Proprietary)** – they may be issued by a controlling organization, but once issued, there is no restriction on the use of the IDs or the descriptive metadata referenced in their definition. Anyone can use the IDs at any time in any workflow without royalty or license.
- **Large** – the ID space must have sufficient capacity to identify all the different assets that might conceivably appear over time.
- **Resolvable** – there is an open and accessible mechanism where the ID can be converted into a description of what it identifies and the reverse, where you can find the ID by describing the identified and the relationship that links the identifier and the identified.

If an identifier fails in any one of these dimensions, then any workflows that depend upon it will eventually fail themselves. Implementing a shared identification system is not without cost, so the stakeholders must take great care to select an identification system that will serve their needs far into the future by being flexible enough to adapt to changing requirements and new applications.

Archives and libraries use a variety of different identifiers within their catalogs and in reference to their collections. Each ID system has its adherents and they can all serve a useful purpose, so long as they are applied carefully and used within their intended domains. Participants in a recent linked open data workshop organized by the FIAF Cataloging and Documentation Commission (CDC), “strongly recommended using unique identifiers, internal ones but also external ones like Wikidata or EIDR identifiers.”

Such identifiers can be grouped into five general categories:
- Location Identifiers
- Classification Systems
- Inventory Indexing Systems
- Statistically Unique Identifiers
- Curated Standard Identifiers

**Location Identifiers**

**Location Identifiers** include Shelf Numbers (for physical assets), File Names and Directory Paths (for digital assets), and URIs (for Internet-accessible assets). “The wanderer in the library is lost without the order that catalogues and shelving systems create.”

---

37 Heftberger, Adelheid. “Linked Open Data for Filmarchives.”
38 Drewry, Raymond, et al. “Content Identification for Audiovisual Archives.”
40 A URI (Uniform Resource Identifier) identifies a resource by name and/or location. URNs (Uniform Resource Name) and URLs (Uniform Resource Locator) are special classes of URIs that are often used with, but not exclusively limited to, the Internet.
locating an asset once it has been identified (e.g., pulling a tape from the shelf once it has been found in the catalog), but are not useful as unique identifiers themselves. They are only meaningful within the context of a particular domain – a library, a vault, a computer, a Web site, etc. – and are not durably linked to the asset being described – if the asset is relocated, then its shelf number/path is changed.

Classification Systems

Classification Systems are most commonly found in libraries and archives, and include the Library of Congress Classification (LCC) system\(^ {42}\) and the archetypal Dewey Decimal Classification system, developed by Melvil Dewey in 1873.\(^ {43}\) Classification systems are at the heart of most library and archive cataloging systems. They were first applied to print works and have since been extended to include audiovisual works and a variety of other asset types. They are “intelligent” identifiers in that meaning may be derived by parsing the classification identifier according to a predefined formula. By their very nature, they include subjectivity in their assignment, so the same work may receive different classifications in different collections even when using the same classification system. Worse yet, different catalogers could assign different classifications within the same catalog. And, since they depend on human intelligence for their generation, they are relatively expensive to apply.

Inventory Indexing Systems

Inventory Indexing Systems consist of a simple identifier assigned by whoever holds or catalogs an asset. They are easy to create, are unique within their domain, and are permanently associated with an asset or a descriptive metadata record (that may then link to the asset itself via a suitable Location Identifier). Inventory Indexing System IDs are popular in accounting systems, media asset management (MAM) systems, title authorities,\(^ {44}\) and publicly searchable databases such as those operated by the British Film Institute, the Complete Index to World Film, and WorldCat.\(^ {45}, 46, 47\) They are most commonly associated with a digital asset

\(^{42}\) Library of Congress. “Library of Congress Classification.”

\(^{43}\) Online Computer Library Center. “Dewey Decimal Classification summaries.”

\(^{44}\) Libraries often manage name and title authorities, where they maintain a single record for each name (person, organization, etc.) and work title in their catalog. These authority records may be assigned a unique identifier, along the lines of a generated primary key in a database table, even if the original source of the authority table was not an electronic data management system. See, for example, the Library of Congress Control Number (LCCN).

\(^{45}\) British Film Institute. “Collections Search: Introduction.”
management (DAM) system and are often a database table primary key. Since they are often simple integers, they can only be guaranteed unique within their domain. The same ID string could be assigned by different parties to reference different things, making it difficult to identify the ID type and its contextualizing domain by simple inspection. Inventory Indexing System IDs can be used as point-to-point identifiers between specific parties, but this does not scale well.

<table>
<thead>
<tr>
<th>Globally Unique</th>
<th>Consistent</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Permanent</td>
<td>? Open</td>
<td>? Resolvable</td>
</tr>
</tbody>
</table>

**Statistically Unique Identifiers**

**Statistically Unique Identifiers** are generated mathematically using a complex formula to return a very large number. They include UUIDs (Universally Unique ID), UMIDs (Unique Material ID), C4 IDs, and file hashes (MD5 and SHA-2). They can be generated by anyone at any time in such a way that they are globally unique for all practical purposes. The most common failing of such systems is also their principal advantage: every time you execute the algorithm it returns a different, unrelated result. This limits their use as shared asset identifiers. File hashes are a special class of statistically unique identifier that use the contents of a digital asset to generate a "message digest" that is unique for each asset – any difference (such as a single pixel changed in a media file) is sufficient difference for an entirely new ID. This means that the ID will always be the same for a given asset no matter who creates it, but such IDs can only be applied to digital assets and there is no association between related assets. For both types of Statistically Unique Identifiers, there is no way to determine what is identified by a particular ID since they are not publicly resolvable IDs.

<table>
<thead>
<tr>
<th>Globally Unique</th>
<th>Consistent</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Permanent</td>
<td>✔ Open</td>
<td>✗ Resolvable</td>
</tr>
</tbody>
</table>

**Curated Standard Identifiers**

**Curated Standard Identifiers** are based on an open, international standard and are administered by an organization that ensures uniqueness.

---

46 Complete Index to World Film. "Film Database Search."
48 May be globally unique within a particular domain (such as WorldCat).
49 ITU-T. "Information technology - Procedures for the operation of object identifier registration authorities: Generation of universally unique identifiers and their use in object identifiers."
50 IEEE. “ST 330:2011 - Unique Material Identifier (UMID).”
51 SMPTE. "SMPTE Standard - Unique Digital Media Identifier (C4 ID).”
52 Rivest, Ronald. "RFC 1321: The MD5 Message-Digest Algorithm.”
53 Lilly, Glenn. “Device for and method of one-way cryptographic hashing.”
54 Statistically unique identifiers are not absolutely guaranteed to be unique in all cases, but the chance that they will repeat is vanishingly small.
and provides ID resolution services. This is the only type of identifier that meets all of the criteria in the Tolkien Rule for identifiers in multi-party applications. These include global identifiers like EIDR (Entertainment ID Registry),\(^{55}\) GTIN (Global Trade Item Number),\(^{56}\) ISAN (International Standard Audiovisual Number),\(^{57}\) and ISBN (International Standard Book Number).\(^{58}\) The limiting factor for such IDs is that they are domain specific. In this case, they are applicable to audiovisual works, consumer products, and books, respectively. The solution is an encompassing identifier family, such as the Digital Object Identifier (ISO Standard 26324).\(^{59}\)\(^{60}\) DOIs are digital identifiers of objects (rather than strictly identifiers of digital objects) and so can be applied to everything from an academic paper (Crossref\(^{61}\)) to a sack of gravel (CIIDRA\(^{62}\)).

- **Globally Unique**
- **Consistent**
- **Permanent**
- **Open**
- **Large**
- **Resolvable**

### Selecting an Identifier System

Why must we agree on a common identifier system? The direct benefit is described by the Protocol Network Effect. This is a direct network effect that arises when a common communications standard is adopted by all of the participants in a network.\(^{63}\) Examples include Ethernet, TCP/IP, telephone exchange system (a.k.a., “dial tone”), VHS (vs. Betamax), and Blu-ray (vs. HD DVD). Imagine trying to communicate on the Internet without IP addresses or domain names. Now imagine trying to operate an efficient multi-institution information exchange without a common way to identify each asset.

Agreeing on a common identification system avoids a specific negative network effect (also known as a “reverse network effect”\(^{64}\)) that would otherwise apply to every transaction on the network. If participants in a communication network do not use a shared identification system, then each party in an exchange must translate the foreign party’s asset identifier into their own system. The nature of the assets held by memory institutions often confounds automated identification schemes that rely on descriptive metadata to determine identity, resulting in the need for manual intervention at each point of exchange. This adds significant overhead,

---

\(^{55}\) EIDR. “About EIDR.”

\(^{56}\) GS1. “Global Trade Item Number (GTIN).”

\(^{57}\) ISO. “Information and documentation -- International Standard Audiovisual Number.”

\(^{58}\) Bowker. “About the ISBN Standard.”

\(^{59}\) International DOI Foundation. “Key Facts on Digital Object Identifier System.”

\(^{60}\) ISO. “Information and documentation -- Digital object identifier system.”

\(^{61}\) Crossref. “About us.”

\(^{62}\) Construction Industry ID Registration Agency (CIIDRA). “Home Page.”


\(^{64}\) Choudary, Sangeet Paul. “Reverse Network Effects.”
complexity, and potential for error to each transaction, which could easily negate all the positive network effects that might otherwise accrue. The initial application of a unique identifier may require manual review, but all subsequent transactions that use the identifier avoid this additional cost and delay, resulting in cumulative savings and enjoying the benefits of the various network effects.

Identifiers are part of an identifier system, which may be self-published (e.g., shelf numbers, classification systems, etc.) or managed by an identifier publisher (e.g., curated identifiers). The authors of “Creating Value with Identifiers in an Open Data World” provide a set of guidelines for identifier publishers to ensure the utility and viability of their identification systems. These guidelines are:

- **provide a reconciliation API** when sharing their own identifiers to allow consumers to match entity names and other characteristics to their identifiers
- **expose documentation** for management of new and existing data frameworks covering the process for assigning identifiers
- **prefer HTTP URLs over other URIs**, ensuring that these resolve to useful metadata about the individual entity
- **ensure that identifiers can reliably be [resolved]** by data consumers and that URL identifiers are created under stable, persistent domain names
- **provide a stable, highly available means of [resolving] identifiers** that they are committed to providing long term
- **should not delete identifiers once in use** so that objects with only historical existence or objects that have been administratively deprecated can continue to be [identified], returning metadata to indicate their state and, where necessary, linking to any succeeding objects
- **avoid using or creating identifier schemes that allow identifiers to be recycled**
- **provide ways for data consumers to track and synchronise changes** to entities that may affect status or identity, e.g., downloadable daily ‘digests’ of changes to identifiers and core metadata, http-based [resolvable] identifier URLs or other synchronization options

---

66 An HTTP URL (HyperText Transfer Protocol Uniform Resource Locator) is a particular type of URI (Uniform Resource Identifier) that can be resolved to content at a particular location via the Internet, typically as a Web page. Not all URIs are resolvable on the Internet.
67 Converting an identifier into the location where descriptive data are stored, in this case via a formula that creates a persistent URL from the identifier.
The ideal identifier for multi-party, global data and process exchange must therefore be unique, permanent, consistent, open, large, and resolvable. The identifier publisher must provide a suitable API, useful documentation, and stable URL resolution. The identifier publisher must also issue permanent IDs that are never recycled and provide a mechanism for users to track and synchronize any system changes.

Few identifier systems can rise to this compound challenge. When the additional constraint that the identifier system must provide IDs that are useful across the myriad asset types held by an audiovisual archive or library, only one viable candidate remains: the Digital Object Identifier (DOI) System from the International DOI Foundation (IDF).

### The Digital Object Identifier (DOI) System

The DOI (Digital Object Identifier)\(^{68}\) System (ANSI/NISO Z39.84 and ISO 26324) was launched in 2000 and now includes over 175 million identified assets and serves over 5 billion ID resolution transactions per year.\(^{69}\) The DOI family of identifiers can be applied to every asset held by an institution from a work in the abstract to a particular physical copy on a shelf.\(^{70}\) Such curated identification facilitates universal search, cross-referencing, academic citation, and inter-archive collaboration while reducing the costs of identification and expanding the efficacy of automation. DOI System identifiers include Crossref (academic papers), DataCite (raw data), EIDR (audiovisual works), ISBN-A (published books), and ARDI (digital rights statements). New identifiers can be added to the DOI System to accommodate the additional asset types cataloged by audiovisual institutions, including posters and other advertising ephemera, un-published written materials, still photographs, props and memorabilia, etc.

The DOI System already provides a universal resolution service, which returns a DOI’s descriptive metadata or the identified object itself. To turn a DOI ID into a resolvable URL, simply add “https://doi.org/” ahead of the ID. For example:

- The Crossref article ID 10.21953/lse.5oi7jyg00b67 can be resolved using [https://doi.org/10.21953/lse.5oi7jyg00b67](https://doi.org/10.21953/lse.5oi7jyg00b67)
- The ARDI rights identifier 10.32102/ardi:1552690256267 can be resolved as [https://doi.org/10.32102/ardi:1552690256267](https://doi.org/10.32102/ardi:1552690256267)

---

\(^{68}\) Digital Object Identifiers are digital identifiers of objects, encompassing physical, digital, and abstract objects, not limited to only identifiers of digital objects.

\(^{69}\) International DOI Foundation. “Factsheet: Key Facts on Digital Object Identifier System.”

\(^{70}\) Gasiorowski-Denis, Elizabeth. “Digital object identifier (DOI) becomes an ISO standard.”
The EIDR Content ID 10.5240/CA2D-7927-3635-DD8D-7D75-U can be resolved using https://doi.org/10.5240/CA2D-7927-3635-DD8D-7D75-U\textsuperscript{71}.

The next evolution of universal resolution will be universal search, whereby core descriptive metadata are turned into the associated DOIs.

The British Film Institute (BFI), the US Library of Congress, and The Media Institute of University College London (UCL) are examples of cultural heritage organizations and academic institutions that are already using DOIs to identify audiovisual assets and facilitate automation.\textsuperscript{72} EIDR IDs are currently cross-referenced with the Internet Video Archive (film and TV trailers), IMDb (descriptive metadata), Rotten Tomatoes (critical and viewer ratings), and well over 200 other types of third-party identifiers outside the DOI System, including memory institutions such as the American Film Institute, Danske Filminstitut, British Film Institute, Svenska Filminstitutet, and Technische Informationsbibliothek.\textsuperscript{73} EIDR and Crossref IDs are also being used in academic paper citations to specifically identify referenced works.\textsuperscript{74}

**The DOI Family of Identifiers**

The following are a representative sample of the identifiers available today within the DOI system that are directly applicable to the assets of an audiovisual memory institution’s collections and catalogs.

**EIDR (Entertainment Identifier Registry)**

EIDR Content IDs reference audiovisual works, the primary assets for an audiovisual library or archive, including linear (traditional movies and TV) and non-linear works (games, interactive TV, etc.); audiovisual, visual-only, and audio-only works; works preserved in analog or digital forms; works recorded to film, tape, disk, or paper; works in their entirety or clips of longer works; different creative versions (cuts or edits); different technical versions (encodings, localizations, etc.); works on their own and assembled into collections; etc.

Mostly, these are commercially-released (actuality, documentary, scripted fiction, theatrical, television, home entertainment, etc.) or professionally-produced works (commercials, music videos, industrials, training films, etc.).

---

\textsuperscript{71} DOIs can also be resolved directly by their respective Registration Agencies in addition to using DOI universal resolution: e.g., https://ui.eidr.org/view/content?id=10.5240/CA2D-7927-3635-DD8D-7D75-U.

\textsuperscript{72} EIDR. “EIDR Members.”

\textsuperscript{73} EIDR. “EIDR Web UI.”

\textsuperscript{74} Viz. de la Pava Velez, Benjamin. “Celluloid Love.”
but an institution may hold additional audiovisual works of historical, cultural, or scientific significance. For example, the US National Film Registry at the Library of Congress includes the *Fuentes Family Home Movies Collection*, the *Reverend Solomon Sir Jones Films*, and the *Zapruder Film*; the *HistoryMakers* is a collection of first-person accounts of African American oral history; and the *Technische Informationsbibliothek* (TIB) maintains an audiovisual archive of scientific experiments and demonstrations – all of which have EIDR IDs.  

All of these different types of audiovisual works in all their different aspects can be individually identified using the EIDR system. EIDR was established in 2010 and the EIDR Content ID registry now contains more than 2 million identified assets with more than 4 million linked alternate identifiers. e.g., 10.5240/CA2D-7927-3635-DD8D-7D75-U

---

**Figure 4. Typical relationships for non-episodic works in the EIDR Content ID registry.**

**ISBNA (Actionable International Standard Book Number)**

ISBN IDs reference “monographic publications” a.k.a., printed books, plus e-books, audiobooks, chapters, and related products made separately available to the public, “irrespective of whether those publications are made available for sale or on a gratis basis.” 10-digit ISBN IDs (e.g., 0-345-27257-9) were introduced in 1970. They have since been superseded by 13-digit ISBN IDs (e.g., 978-0-345-27257-7), first introduced in 2004 with the

---

75 Library of Congress. “Complete National Film Registry Listing.”
77 TIB. “TIB AV-Portal.”
78 EIDR. “About EIDR.”
79 For Episodic works, the Collection layer contains additional Series and Season elements (using North American parlance) while the Abstraction layer contains the Episodes.
80 ISO. “ISO 2108:2017.”

Figure 5. The International Article Number (EAN) family of barcode identifiers.

Crossref
Crossref IDs reference scholarly communications, including “journals and journal articles, books and book chapters, conference proceedings and papers, reports, working papers, standards, dissertations, datasets, and preprints.” Crossref was established in 2000 and now supports over 12,000 members in 114 countries around the world and has issued over 100 million Crossref IDs. In addition to providing DOIs for traditional scholarly communications, Crossref also offers DOIs to reference “Event Data” for Internet-based communications: “Events are comments, links, shares, bookmarks, references, etc. … When someone links their data online, or mentions research on, for example, Twitter, Wikipedia, or Reddit, we capture that Event and make it available for anyone to use in their own way.”

DataCite
DataCite IDs reference research datasets and other research objects identified by data owners, stewards, and archives “to help the research

---

83 GS1. “The GS1 Barcodes.”
85 Crossref. “Help depositing metadata.”
86 Airiti DOI, JaLC (Japan Link Center), and KISTI (Korea Institute of Science and Technology Information) offer similar local market academic writing and data set identification services in Taiwan, Japan, and Korea, respectively. CNKI (China National Knowledge Infrastructure) and ISTIC (The Institute of Scientific and Technical Information in China) offer China-based services.
87 Crossref. “About us.”
89 Crossref. “Event Data.”
community locate, identify, and cite research data with confidence.”

DataCite’s members include data centers, libraries, government agencies, and research universities in more than 20 different countries.

**mEDRA**

mEDRA provides DOI registration services to publishers, academic institutions, and their intermediaries in Italy, the EU, and globally, and provides ISBN-A services. mEDRA also offers the ARDI service, which provides DOIs for rights declarations and copyright declarations.

**People Identifiers**

ISNI, the International Standard Name Identifier, references “contributors to creative works and those active in their distribution, including researchers, inventors, writers, artists, visual creators, performers, producers, publishers, aggregators, and more.” Its application is limited, however, because it is literally just that – a “name” identifier. ISNI IDs disambiguate people’s names. If people use different names during the course of their life or professional career, then each name could have a different ISNI ID. There is no guarantee of a single identifier applicable to all the works of a given individual. Various commercial and non-commercial IDs exist in this space (from vendors including Gracenote, IMDb, ORCHID, Red Bee Media, etc.), but they do not satisfy all of the requirements for a universal identifier outlined above. A DOI-based person identifier that will meet these needs is under active discussion within the Media and Entertainment industry.

**Universal Search: Something New Under the Sun?**

What has been will be again,
what has been done will be done again;
there is nothing new under the sun.

Multi-institution search is not a new thing. Many centralized, aggregated, and federated search services have been developed in the past and many are operating now. They differ in how they build their index, in their scope, and in their use of universal identification. A representative sample of currently active multi-institution search initiatives would include:

---

90 DataCite. “Our Mission.”
91 DataCite. “Members.”
92 mEDRA, “What is the ARDI service?”
93 mEDRA, “Who we are.”
95 ISNI International Agency. “International Standard Name Identifier (ISO 27729).”
96 Eccles. 1:9 New International Version.
- **ArchiveGrid:** Multi-archive search service that extracts selected MARC records from WorldCat and periodically scrapes registered Web sites that provide EAD XML finding aids.\(^97, 98\)
- **Archives Portal Europe:** A search portal that “provides access to information on archival material from different European countries as well as information on archival institutions throughout the continent.”\(^99\)
- **Digital Public Library of America:** A search portal that “aggregates metadata ... and thumbnails for ... photographs, manuscripts, books, sounds, moving images, and more from libraries, archives, and museums across the United States.”\(^100\)
- **European Holocaust Research Infrastructure's (EHRI) Portal:** A catalog of electronic descriptions of the Holocaust-related assets held by participating institutions.\(^101\)
- **Europeana:** A European search portal for “artworks, artefacts, books, films and music from European museums, galleries, libraries and archives” based on metadata submissions from contributing institutions and aggregators.\(^102\)
- **The National Archives (UK):** Search records of assets submitted by participating UK archives.\(^103\)
- **WorldCat:** – Global aggregated library catalog search from an index of MARC records submitted by participating institutions.\(^104\)

A contributor to this paper summarized the state of cross-archive search in the UK, which contains a typical mix of topic-specific, geography-specific, and national archive services:\(^105\)

> It’s a rather fragmented picture, for a relatively small country. Which obviously makes it difficult for users to know where to search. And of course it makes it challenging for institutions to know where to share their catalogues, too. .. It is difficult enough to get any cataloguing done, without having to reformat / resubmit the information to accommodate different systems.

A universal identifier system and search service that satisfies the Tolkien Rule can also help here – not by replacing the various portals and aggregators, but by making it easier to submit material to them, easier for them to gather materials on their own, and easier for users to link to and discover assets of interest.

---

\(^{97}\) OCLC. “ArchiveGrid.”
\(^{98}\) Library of Congress. “<ead>Encoded Archival Description.”
\(^{100}\) DPLA. “Home Page.”
\(^{101}\) EHRI Consortium. “Home Page.”
\(^{102}\) Europeana. “Europeana Collections.”
\(^{103}\) National Archives, The. “Home Page.”
\(^{104}\) OCLC. “Home Page.”
\(^{105}\) “RE: [AMIA-L] Multi-Archive Search Services.”
The European Film Gateway (EFG) is an excellent example of a multi-intuition aggregated search service. It indexes over 600,000 assets, including "photos, posters, programmes, periodicals, censorship documents, rare feature and documentary films, newsreels and other materials" contributed by 26 different European film archives and cinemathques. It is a manually-curated collection built from 2008-2011, with additional materials related to World War I added from 2012-2014. It has many features required for a successful universal search service, but also several key limitations:

- **Centralised versus decentralised**: EFG offers a centralised portal, with collection owners adding records to the central store, [while the proposed universal search] is a model for aggregating in a decentralised form.
- **Active/manual versus passive/automatic**: EFG is created by collection owners consciously and actively uploading metadata and content, [while the proposed universal search] works in the background, automatically created on the fly by systems.
- **Iterative updates versus dynamic updates**: EFG updates in iterations when contributors undertake their work, [while the proposed universal search] updates dynamically as archives create data in their systems.
- **Selected/curated versus self-selecting/un-curated**: EFG offers a selection of digital collections, [while the proposed universal search] is all collections, analogue and digital, not a selection.

![Figure 6. Comparison of key characteristics of EFG to the proposed Universal Search.](image)

---

106 EFG. “About.”
107 EFG. “About the European Film Gateway.”
108 McConnachie, Stephen. “I’ll show you...”
Including globally-unique, curated identifiers as a core component of universal search offers several key advantages and creates new opportunities not present in earlier, multi-institution search services:\textsuperscript{109}

- **Unambiguous identification of works:** no room for confusion because of matching titles, dates, names. The deduplication and uniqueness has been achieved in the [ID] Registry.
- **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).\textsuperscript{110}
- **Automation of comparison:** the unique identifier allows applications to query APIs\textsuperscript{111} and establish very quickly and easily what is held by the organisation for this work. ... [T]he unique ID is the fuel which makes it run quickly and without anomalous results.

### Citation and Cross-Reference

Proper source citation is the foundation of scholarly writing, research, and academic integrity. Bibliographic reference is “the connective tissue of scholarship.”\textsuperscript{112}

Academic freedom can flourish only in a community of scholars which recognizes that intellectual integrity, with its accompanying rights and responsibilities, lies at the heart of its mission. Observing basic honesty in one’s work, words, ideas, and actions is a principle to which all members of the community are required to subscribe.\textsuperscript{113}

According to Geoffrey Bilder, Crossref’s Director of Technology & Research, “Citation is not just about credit and rights, but is about pointing at evidence and, increasingly, about ensuring that research is reproducible. In this latter case, it is becoming even more important that the resolution of the citation identifier be precise and machine actionable.”\textsuperscript{114} There are a number of different citation standards in common use, including APA, MLA, Chicago, Turabian, and IEEE.\textsuperscript{115} They each rely on manual interpretation and are therefore subject to variation and inconsistent resolution. Using DOIs in

---

\textsuperscript{109} McConnachie, Stephen. “I’ll show you...”
\textsuperscript{110} ID cross-reference, such as provided via the EIDR registry, extends the common machine processable identifier by referencing the domain-specific identifiers (“Alt IDs”) unique to each participating institution, which can then be used for local resolution.
\textsuperscript{111} Application Programming Interface: A standard mechanism for machine-to-machine communications and process automation.
\textsuperscript{112} Brian Cantwell Smith as quoted by Thompson, Henry S. “Understanding URI Ecosystems.”
\textsuperscript{113} Princeton University Graduate School. “Academic and Research Integrity.”
\textsuperscript{114} Bilder, Geoffrey, “RE: DOI for universal identification.”
\textsuperscript{115} Wiggins, LaMonica, et al. “Citation Styles.”
academic citation provides an unambiguous reference to the source material. The DOI citations also work in the reverse direction to provide cited-by references, indicating where and how often each piece of source material is referenced by third parties.

For example, the master’s thesis “Celluloid love: audiences and representations of romantic love in late capitalism” by Benjamin de la Pava Velez includes 111 DOI references in its 42-page bibliography. The DOI citations come from two ID Registries: Crossref IDs for academic papers and EIDR IDs for motion pictures.116 “Celluloid Love” is itself identified by Crossref DOI 10.21953/lse.5oi7jyg00b67, which the DOI resolution service directs to http://etheses.lse.ac.uk/3602/, a London School of Economics Web site where the paper is summarized and can be downloaded.117

Within a memory institution, production materials, manuscript collections, photographs, the screenplay, and the motion picture soundtrack would typically add the in-house work identifier for the motion picture to each record as a related work entry. This cross-reference establishes the relationship between the different assets, but in practice is still heavily dependent on a human readable label.

To facilitate cross-reference among related works in academic writing, Crossref provides an automated cited-by tracking service for authors, leveraging the work’s assigned DOI to identify where each work is referenced in another’s research.118 To facilitate cross-reference among related identifiers, EIDR provides an alternate identifier service where EIDR records contain applicable identifiers from external systems. As a result, there are more alternate identifiers in the EIDR Registry than assigned EIDR Content IDs. See, for example, the EIDR Abstraction record for The Polar Express (2004), which contains more than 50 alternate identifiers including more than 40 that hyperlink directly to third party records related to the film.119 Memory institutions can adopt similar practices by using each object’s DOI as a cross-reference identifier within the related catalog entries, establishing an unambiguous, durable, machine-readable and actionable link between the related assets.

---

116 de la Pava Velez, Benjamin. “Celluloid Love.”
117 London School of Economics. “Celluloid Love.”
118 Crossref. “Cited-by.”
119 See https://ui.eidr.org/view/content?id=10.5240/0E8E-C250-E484-9794-A9F2-0
**Linked Open Data**
Digital cataloging is now the norm for memory institutions. This information is valuable in itself and, as discussed above, it is more valuable when combined with data from other sources. Linked Data (or Linked Open Data, if the data is freely readable and usable by anyone) is the dominant way of doing this. “Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods.”\(^{120}\) Tim Berners-Lee’s four principles of Linked Data are:\(^{121}\)

1. Use URIs as names [identifiers] for things
2. Use HTTP URIs so that people can look up [resolve] those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)\(^{122}\)
4. Include links to other URIs [cross references], so that [people] can discover more things.

That may seem like a lot to take on all at once, but resolvable universal identifiers – like DOI – simplify this task significantly:

1. A DOI is an identifier, and is trivially converted to a URI by prefixing it with “https://doi.org/”
2. DOIs are inherently resolvable
3. Resolving a DOI returns metadata about the underlying referent. RDF support is emerging, but other machine-readable forms such as XML and JSON\(^{123}\) are already available
4. External identifiers in a DOI system can be represented both as a bare identifier and a resolvable one, e.g. 10.3203/IWF/E-1754 and https://doi.org/10.3203/IWF/E-1754, or 150780323 and http://collections-search.bfi.org.uk/web/Details/ChoiceFilmWorks/150780323

Both industry and memory institutions have started exploring and implementing Linked Data systems.\(^{124}\) The Library of Congress’ BIBFRAME (Bibliographic Framework) was established to facilitate the transition from MARC 21 catalogs to Linked Data. It “provides a foundation for the future of bibliographic description, both on the web, and in the broader networked world that is grounded in Linked Data techniques.”\(^{125}\) For more information

---

\(^{120}\) Linked Data. “Linked Data - Connect Distributed Data across the Web.”
\(^{121}\) Berners-Lee, Tim. “Linked Data.”
\(^{122}\) RDF: Resource Description Framework; SPARQL: SPARQL Protocol and RDF Query Language. RDF is a framework for describing Web resources, while SPARQL is a query language for RDF resources.
\(^{123}\) XML: Extensible Markup Language; JSON: JavaScript Object Notation. Both are used on the Web to structure data for machine parsing.
\(^{124}\) McCrae, John P., Editor. “Linked Open Data Cloud.”
\(^{125}\) Library of Congress. “Bibliographic Framework Initiative.”
on progress within the film and television industry, see “A Creative Works Ontology for the Film and Television Industry,”\textsuperscript{126} and for the heritage sector, see “Linked Open Data for Film Archives.”\textsuperscript{127}

At its core, Linked Data is about using resolvable identifiers to connect data sources and using standards to communicate the data. DOI provides infrastructure for this second part as well, through its support of HTTP content negotiation. Using content negotiation, a client can ask for results in different formats, according to its needs. CrossRef, DataCite, and mEDRA provide both general formats and domain-specific formats.\textsuperscript{128} EIDR supports several different resolution formats as well.\textsuperscript{129}

Resolvable, unique identifiers are the foundation of any Linked Data system. Once an institution’s collections and catalogs have been described and identified, publishing them as Linked Data is a relatively straightforward process. Linked Data is not an all or nothing choice. Institutions can approach it at their own pace, starting with the parts of it that are of most value to them and their communities of users.

**Bringing It All Together**
All this change does not come for free, but it is necessary. As institutions themselves change by adapting to the evolving nature of their collections and the ways in which they are accessed, the way they manage their assets and user interactions must change as well.

Breaking old paradigms ... by allowing a new universe of information into the decorous space of the library, [is] useless or even dangerous unless there [is] a new paradigm to take its place, a new vision of what these expanded worlds meant. Without this, those who had once felt at home in the world would simply be stranded in a pathless sea of information.\textsuperscript{130}

The first, and simplest, step is to take advantage of what is available now.
- Memory institutions can obtain EIDR IDs for audiovisual works without charge, for both new acquisitions and for current catalog items.\textsuperscript{131} It is

\textsuperscript{126} Motion Picture Laboratories. “A Creative Works Ontology for the Film and Television Industry.”
\textsuperscript{127} Heftberger, Adelheid. “Linked Open Data for Filmarchives.”
\textsuperscript{128} CrossCite. “Supported Content Types.”
\textsuperscript{129} EIDR. “EIDR and the DOI Proxy.”
\textsuperscript{131} In exchange for assistance promulgating the EIDR standard. (See “Beyond Entertainment – EIDR Welcomes UAM’s (Universities, Archives, and Museums) to Participate in EIDR.” https://eidr.org/uam.)
also good practice to ask producers, distributors, donors, and other sources for any EIDR IDs they have for new acquisitions.

- Book suppliers should be asked to provide ISBN-A IDs for all new acquisitions and academic publishers should be asked to provide Crossref or similar DOI IDs for all published papers.
- Memory institutions that produce publications or new data sets should register them with Crossref, DataCite, or an appropriate local market DOI Registration Agency.
- Any sacks of gravel in an archive’s collections should be registered with CIIDRA, and so on.

Ideally, such registrations will be included as standard practice when new assets are acquired, when existing assets are cataloged, and when any existing catalog entries are updated. To support this:

- Add the unique identification and the supporting practices to existing cataloging curriculum and future cataloging training programs.
- Work with affiliated academic institutions to ensure that the importance of unique identification is communicated to all who could benefit from it and work in fields that require it.

If institutions identify DOI coverage gaps, then they should be filled. First, by encouraging existing DOI registries to expand descriptive metadata and best practices to accommodate the particular needs of memory institutions. Next, by coming together to expand DOI coverage for asset types not presently addressed by an existing DOI registry, such as people, production artifacts, and advertising ephemera.

As more and more of an institution’s collections and catalogs are uniquely identified, cross-institution cooperation becomes more streamlined

- It will be increasingly easy to establish multi-institution search and discovery services, and to cross-link and integrate those that already exist, eventually leading to universal search.
- Shared catalogs can be published via Linked Data, facilitating new uses and workflows that have yet to be conceived.

The participating institutions will not have to do all the work here. Once data becomes available, self-forming communities of interest will start to make use of it to address their particular needs and the benefits will be shared by all. As an example, EIDR’s original target workflows centered on digital distribution, but now the same set of data and links is used by archives and researchers as well.
Conclusion
According to Metcalf’s Law, an interconnected network of memory institutions has a value greater than the simple sum of its participants and, as a result, the value of each institution increases as a result of its participation in the network. The value delivered to institutional stakeholders increases at an even greater rate due to the social network effect and Reed’s Law. Once the network starts to grow, additional institutional members and participating users will join to share in the benefits derived from the Marketplace Network Effect.

In addition to the beneficial network effects, a robust and durable system of universal identification applied across all assets held or cataloged by memory institutions can lead to further benefits through automation, universal search, resolution, academic citation, cross-reference, and Linked Data.

Using DOI as a universal identification system for memory institutions will not mean the end of cataloging, since identification without descriptive metadata is of limited use. When metadata and identification are combined, they are tremendously powerful. The goals of robust identification are to connect sources together and to use those connections to improve the discoverability of information about the identified works. Every institution has different information and supporting material for a work – reviews, transcripts of local-language dubbing, national film festival information, publicity stills, posters, academic research, and so on. DOIs will help extend cataloging and open up new opportunities for connection, collaboration, access, and sales.

The initial groundwork has been laid, now the audiovisual library and archive community must come together to realize these collective benefits.
Appendix: Common Audiovisual Memory Institution Asset Classes

The collections and catalogs at audiovisual libraries and archives contain far more than just audiovisual works. There are myriad related and collateral assets, each with its own collection, preservation, and cataloging issues – and each with its own identification needs.

- **Audiovisual Works**
  - **Work Types**
    - Theatrical motion pictures and shorts
    - Television programs, one-off and series
    - Home entertainment and direct-to-video
    - Born digital Web originals
    - Actualities
    - Industrials
    - Educational programs
    - Commercials
    - Experimental films
    - Propaganda
    - Outtakes, B-roll, and stock footage
    - Home movies, both film and video
    - User generated content (UGC)
    - Etc.
  - **Media**
    - Film
    - Tape, analog and digital
    - Disk, magnetic and optical
    - Paper\(^{132}\)
- **Advertising Ephemera**
  - EPKs
  - Featurettes
  - Festival Catalogs
  - Lobby Cards
  - Marketing Assets (sizzle reel, infographics, dataviz, etc.)
  - Portrait Stills
  - Posters
  - Pressbooks
  - Publicity Stills
  - Social Media Assets (a Twitter feed, Facebook page, etc.)

---

\(^{132}\) In the early days of cinema, motion pictures did not qualify for copyright protection so many producers printed their films on paper, frame by frame, and registered that for copyright. In some cases, these paper copyright filings now represent the only extant original copy of the work.
Institutions hold two types of legal documents: those collected for their cultural, historical, or commercial value and held as part of the institution’s collections; and those that govern access to an institution’s collections and catalogs, since not everything held by an institution is in the public domain or free of donor restrictions.

Technically, people are not held by a library or archive, but they are important to documenting an institution’s collections and annotating the interconnection between different works in an institution’s catalogs.
About the Authors

**Richard W. Kroon** is the Director of Engineering for the Entertainment Identifier Registry Association (EIDR). Prior to EIDR, he held senior positions with the Motion Picture Association of America (MPAA), Technicolor, MovieLabs, and Southbay Motion Picture Technologies. In addition to being an award-winning videographer, Mr. Kroon is the author of numerous technical papers and several books, including *A/V A to Z* and *3D A-to-Z*. Mr. Kroon has a BS in Computer Science from the University of Southern California (USC), an MBA from Auburn University, an MA in Media and Communications Psychology from Touro University Worldwide (TUW), and post-graduate certificates in Film and Television from UCLA Extension’s Entertainment Studies.

**Raymond Drewry** is Principal Scientist at MovieLabs and Chief Architect at EIDR. His current interests are identifiers and identification; metadata and linked data; and distributed and federated systems, especially in television, film and the broader entertainment industry. He has held management and technical roles at Liberate Technologies, Sybase, Digital Equipment Corporation, and Microsoft, and is Vice Chair of the DOI Foundation. He has previously worked in computer graphics, IP-connected television, and performance art using semi-autonomous heavy machinery, and has a BA in Classics (Latin) and Computer Science from Yale University.

**Andrea Leigh** is the Moving Image Processing Unit Head at the Library of Congress Packard Campus for Audio Visual Conservation. Previously, she was a cataloger at the UCLA Film & Television Archive and held positions at the UCLA Library, Cal State University at Northridge, Walt Disney Company, and Creative Artists Agency (CAA). She has a BA in Theatre Arts and a Master of Library & Information Science (MLIS) in Moving Image Archive Studies and from UCLA.
Stephen McConnachie is the Head of Data and Digital Preservation at the British Film Institute (BFI), leading the Collections Systems, Information Specialist and Documentation Editor teams, delivering data and systems objectives for the BFI National Archive and BFI Reuben Library. His key responsibilities include the Digital Preservation Infrastructure, Collections Information Database, the BFI Filmography, and the Contribute crowdsourcing platform. He has 20 years’ experience leading and managing data, systems, and documentation teams and projects, defining strategies and workflows for data and digital preservation, procuring and integrating collections systems, and developing Web applications to surface collections and data in innovative ways. He holds a BA in English and History from Northumbria University and an MA in Film and Television Studies from University of Warwick.

Bibliography


Sturdy, Christopher B. and Richard Mooney. “Bird communication: Two voices are better than one.” Current Biology, Vol. 10, Issue 18 (September 2000), R634-R636. https://doi.org/10.1016/S0960-9822(00)00661-8


https://doi.org/10.5281/zenodo.2551045


