EFFICIENT WORKFLOWS FOR AUTOMATED RICH METADATA PRODUCTION

Richard Kirk, Red Bee Media, UK

Ben Schofield, EIDR, UK

Don Dulchinos, EIDR, US

ABSTRACT

Rich metadata has become an essential component in media distribution, primarily to improve discovery and navigation for the consumer across an increasing number of platforms and services. Metadata production costs need to be reduced to maintain reach across these new platforms and create an efficient bridge between broadcast and digital services. Raw materials are drawn from many sources, which often use their own formats and proprietary identifiers, making interoperability error prone, expensive, and difficult to scale. This paper covers new approaches to automated matching of metadata based on three features of the EIDR unique identifier: resolvability, explicit support for identifiers from other systems, and the availability of a web-based API. It does this in the context of the day-to-day practical experience of delivering multi-lingual rich metadata for many of the leading broadcasters and digital platforms in Europe such as the BBC, ITV, UPC, TV4, and BBC iPlayer and the workflows used for US film and TV distribution. Red Bee Media combined library matching with near real-time updates from linear broadcast to improve editorial efficiency and the quality of metadata workflows. This implementation points the way to new best practices for metadata that enable and simplify next-generation applications, such as contextual linking and unified audience measurement

INTRODUCTION

How will you find your next TV programme? Chances are that you will browse a guide on your set-top box, look at images and trailers or read additional text. You may read a newspaper review and search on Google for more information or visit a specialist web resource such as IMDB.

When you watch the programme it is likely that you use a variety of devices, playing back content from TVs, set-top boxes, consoles or mobile devices and expect the experience to be consistent and the information accurate and up to date.

The cost of manual metadata production is high, major broadcasters such as the BBC invest heavily in the editorial enhancement of their metadata and want to make sure it is used efficiently across all platforms.

As the digital supply chain expands to supply transcoded content to new devices, there is a profusion of content formats and deliveries. The legacy manual supply chain processes cannot cope with the new volumes or deliver at a low enough cost.

This paper covers new approaches to automated matching of metadata based on the EIDR unique identifier. It describes how ID's are resolved, the use of identifiers from other systems, and how a web-based API can be integrated with other tools into existing workflows. It does this in the context of the day-to-day practical experience of delivering multi-lingual rich metadata for many of the leading broadcasters and digital platforms in Europe such as the BBC, ITV, UPC, TV4 and BBC iPlayer; it also discusses some of the workflows used for US film and TV distribution.

UNIQUE CONTENT ID'S

Telephone numbers are universal, with the right international country code a caller from one country can connect directly to someone in another country without having to understand which underlying systems are in place and how they work. If only the same was true of content ID's, there is no "metadata dial-tone" to track content through the supply chain, make recommendations to viewers or measure consumption.

Many content ID's today are specific to a particular system or domain and can't be used out of that context. ID's often contain information that is human-readable, lending themselves to direct editing without any checks. Content may be duplicated with many ID's for the same asset, or worse many assets for the same ID. This leads to a great deal of wasted effort in trying to establish the precise version of a piece of content. While manual trafficking processes worked fine for premium high-value assets in simple distribution chains such as cinemas, they are very expensive and unsustainable for today's digital distribution.

The ideal unique ID is:

- Abstracted - requiring a check on a central database, unable to be edited by (well-meaning) humans
- Persistent – consistent over time, irrespective of owner or industry body
- Hierarchical – allows content to be grouped and variants tracked
- Standards backed – industry endorsement, based on existing standards where possible and in turn referenced by other emerging standards
- Coverage - able to generate unique identifiers for all types of valid audiovisual content types including the emerging, high-volume categories such as clips.
- Flexible - support a wide variety of relationships and hierarchies that can exist between assets.
- Extensible - readily extensible to accommodate new types of assets that may emerge as the industry evolves.
- Scalable - handle very large volumes of registrations and perform reliably in order to be integrated into automated workflows.
- Cost-effective - make economic sense for the entire ecosystem, even at very high-volume scale which allows large archive collections and libraries to be added without major financial impact
- Interoperable - interoperate with and support other existing IDs and databases.
- Value-added services - support innovation by focusing first on a unique namespace and identifying metadata, enabling commercial providers to offer new and existing value-added metadata and other services on top of the system without restriction.
- Accessible - broadly available and provide easy search and query access to all commercial ecosystem participants irrespective of their commercial scale.

EIDR ID's have been built to fulfill these aims.

ENTERTAINMENT ID REGISTRY

The US studios tried a number of ID approaches and schemes with limited success and commissioned MovieLabs to perform a major global study of existing content ID schemes before creating the EIDR standard. One of the key aims was to make it usable by content owners, distribution platforms, broadcasters and other entities in the value chain without prejudicing their individual business models

The founding members in August 2010 were MovieLabs, CableLabs, Comcast and Rovi together with the major studios. Their mission was to drive out inefficiencies in the supply chain, while enabling new business models and workflow processes to support the increasing opportunities and complexities in digital distribution of content. EIDR provides not just standalone IDs, but also linked sets of IDs that can describe and link all of a works versions, encodings, and related content with the inclusion of identifiers from other systems, allowing interoperability across those systems.

By the summer of 2014, the EIDR consortium numbers close to 70 member companies from the US, United Kingdom, France, Germany, Switzerland, The Netherlands and Japan. Some key members, including Microsoft, Google and Vudu, are working with studios to define a standardized architecture for digital content distribution and retailing. These industry-wide efforts to standardize on taxonomies, APIs and unique numbering will help unlock the potential of these currently loosely-coupled data sets.

As an application of the Digital Object Identifier (DOI) standard, EIDR is built to be interoperable with other identifier systems such as ISAN, ISRC and Ad-ID, and aims to provide a core service to vendors who wish to develop offerings such as rights management and extended metadata services. The EIDR data model can be readily extended to cover new and emerging objects and relationships as the industry evolves over time.

HOW IT WORKS

The registry assigns a unique universal identifier for all registered assets. EIDR is an opaque ID with all information about the registered asset stored in the central registry. Its structure consists of a standard registry prefix, the unique suffix for each asset and a check digit. The suffix of an asset ID is of the form XXXX-XXXX-XXXX-XXXX-XXXX-C, where X is a hexadecimal digit and C is the ISO 7064 Mod 37, 36 check character [1].

Standard prefix for EIDR registry	Unique suffix for each asset	Check digit

10.5240/XXXX-XXXX-XXXX-XXXX-XXXX-C

EIDR is purely functional without any implication of ownership, making it persistent enough to remain the same despite any change in control or ownership of the underlying asset. The Registry requires the minimal metadata necessary to guarantee uniqueness for the full range of asset types.
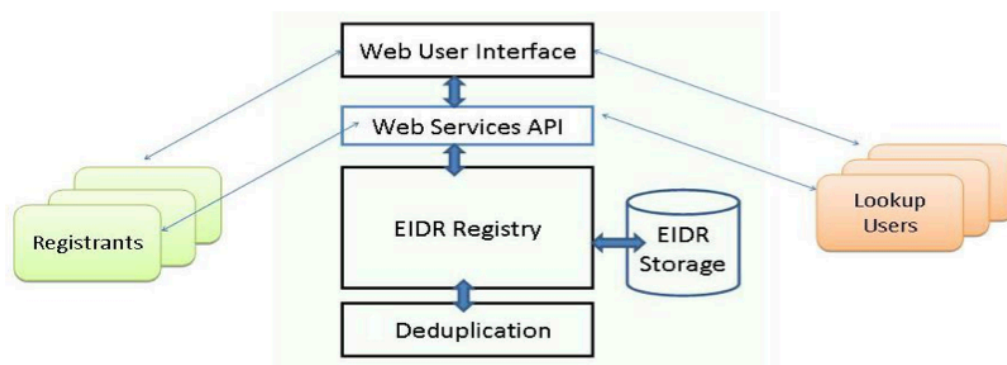
Figure 1: Entertainment Identifier Registry diagram

The registry receives and processes registration requests from registrants. Users and applications can look-up IDs, match their own records against the EIDR database and search the registry. Registrants and lookup users can use the web user interface or web services API to interact with the registry. There are a variety of tools and SDKs to help engineer the integration and an active community of operations experts drawn from the membership that agree and document best practice and drive adoption in their organisations.

Registrants Submit Assets to the Registry

A registrant submits objects for matching and registration along with core metadata and information such as the type of object and relationship to other objects. The registrant could be a content owner, an aggregator, post-production house or any other entity authorized to register objects.

Registry Assigns EIDR After Verifying Uniqueness

EIDR uses a sophisticated de-duplication system to ensure that the object submitted to the registry has not already been registered while allowing the registration of similar and related objects. The centralized registry structure guarantees the uniqueness of objects being registered. If no duplicate object exists, the registry generates an EIDR for the object and stores the new EIDR and the corresponding metadata in the registry. If a duplicate of the new object already exists, the existing ID is returned to the registrant.

Lookup users can search the Registry

The Registry provides both a web-based UI and web-services APIs to enable a user or application to search the registry using various search criteria. Developers can use the APIs to integrate the registry features with their applications and automated workflows.
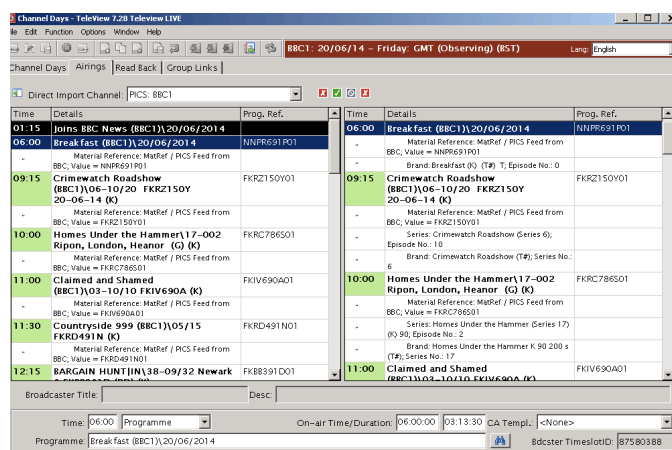
RED BEE MEDIA AND EIDR

At Red Bee Media, our highly structured and functional metadata has long underpinned the UK and European media landscape with a database of over 8 million uniquely identified programmes, series and contributors.

Our core Teleview editorial system:

- automatically processes data from broadcaster scheduling systems and other feeds
- manages complex editorial workflows to ensure EPGs meet publishing deadlines,
- integrates with play-out systems at TX to maintain highly accurate feeds for PVR recording and VOD schedules
- delivers multiple feeds to systems with near-real time updates

In the early days Teleview's core data structure was fully mapped against the EBU TVAnytime standard, we worked closely with the BBC on their TVA implementation for BBC iPlayer and Youview and participated in the early technical discussions on EIDR. It was a natural extension of this approach for Red Bee Media to become a full member and start the EIDR data mapping and integration of our processes.

Although EIDRs initial focus was on digital supply chain efficiency there is increasing interest in rich metadata to help consumers find the best content. Better connected data sets that span multiple markets are the key to enabling the next generation of media apps and Content Discovery services. Red Bee see EIDR as a key element of these solutions, enabling us to help our broadcaster, platform and other industry partners launch the exciting next generation of media based apps and services."

Getting started

The first steps involved the core technologists and leading editors, with a technical review of the documentation at http://www.eidr.org that covers both the technology and operational best practice. This led to some basic manual matching of small volumes of assets sampled from the database and direct read access to the system through https://ui.eidr.org. After some practice registrations through a "sandbox" registry the editors are cleared by the EIDR onboarding team for write access to the main production database.

Bulk matching of "simple" assets

For the first phase of our matching process we chose assets that are simple to understand and easy to resolve any questions about. Red Bee works in over 30 languages so we chose only English titled movies to avoid any language complications, movies are easier as they are mainly stand-alone assets with no or little hierarchical structure.

We implemented a semi-automated phased matching of approximately 59,000 of these movies and achieved a match on around 42,000. This match rate showed that the assets we'd selected were largely popular US or international movies.

The minimum data set to successfully determine a match or registration is:

- Title and Title Language
- Original Languages and Manifestations
- Release Date
- Approximate Length
- Country of Origin
- Participants (must have 1 AssociatedOrg or Director or 4 Actors, more are better)
    - AssociatedOrg and Role (Production company, distributor etc)

- Directors (up to 2)
- Actors (up to 4)
  - Good to also include (Alternate Title, Alternate ID (IMDb, ISAN, Internet Video Archive, Proprietary IDs))

Bulk matching of other assets

Next we are moving on to TV series and movies in all languages. The language referred to here is the language of the metadata associated with an asset, many of the assets in our dataset will already be registered. There will also be assets that only appear on our database in one, non-English language. E.g. *Forbrydelsen (The Killing)* when it first appeared in Denmark.

Transliterated titles also present a challenge when registering content. If we register an asset with a transliterated title and another company registers the same asset with the original title, this will cause a duplicate entry. . EIDR best practices strongly encourage the addition of titles in non-original languages when they are available. For example, the EIDR record for Sophie's Choice has 10 alternate titles and 12 alternate IDs. See https://ui.eidr.org/view/content?id=10.5240/CA23-7B6E-E6CD-07D7-68F2-E

For this reason we are planning carefully which of these assets to register and which to leave until there is a compelling reason to tackle a more complex set of data.

What we won't match

We have focused our efforts to date on the most valuable class of metadata assets, those that often scheduled on linear channels or on-demand services. Premium movies and hit TV series are our first target. The content that is only shown once is less attractive for us to match (unless specifically requested by one of our customers). This includes live events (music, news & sport) and one-off documentaries.

Registration

Now that we've got to the point of matching against what is already on the EIDR registry, we will also register assets from our database that don't yet exist on EIDR. Red Bee will "own" those EIDR records, but the content owner can take control of editing those records through a simple process with the EIDR operations team. As back catalogues are bulk registered and new content is born digital with an EIDR ID we expect that root registration effort to reduce and for our teams to focus on more granular registration of variants.

Making the internal business case

As in any service organisation diverting resources away from current business requires a robust case. For Red Bee this covered the annual membership fee, editorial and technical resource, it helps to have a client or potential client that currently uses EIDR or could benefit from its deployment

We have over 160 editors who annually spend many thousands of man-hours moderating the metadata for clients, even minor efficiencies can deliver significant savings. We were able to break out individual client workflows and consider the potential impact of better matching, tracking of variants and inclusion of third party references.

Key technical functionality

For Red Bee there are some key attributes of EIDR that we have been able to exploit in our workflows

- Resolvability – the ability to check on the attribution of a piece of content with a single call, many of our customers find it difficult to deliver bulk data but can easily make an external call to verify an item
- Explicit support for identifiers from other systems – we absorb content from a variety of feeds and sources, the retention of a range of different identifiers such as IMBD or in-house production numbers makes disambiguation fast and efficient
- Web-based API – standard based approach allows us to quickly link systems together without major development or ongoing code maintenance

Short term gains

The main business driver for Red Bee adoption of EIDR is to help our broadcast partners' track content for rights, royalties and analytics. Our industry partners for some other data sets, like trailers, have already

adopted EIDR, we are seeing short-term benefits in the partially, ready-matched asset library. The major US film and TV studios are starting to provide VoD avails and TV schedules with EIDR IDs as the key reference, significantly reducing the amount of manual verification and matching that takes place today due to poor quality data.

New applications of EIDR

Red Bee carry out localisation and transcoding services in the UK for the European distribution of major US studio content. The EIDR ID enables our teams to accurately identify the core content and combine metadata from a number of disparate studio systems and external sources. Although the internal studio workflows rely on multiple proprietary ID's held on in-house systems, these are now added by the studio to the EIDR web registry as Alternate IDs at key workflow stages. EIDR is used as a universal lookup to return the corresponding EIDR ID from the in-house ID reducing the manual trafficking effort normally expected in this process. EIDR can be adopted incrementally without forcing major system changes. Red Bee can register new EIDR IDs for the editorial variants and digital packaging commissioned by the studios and ensure they are included within the delivery for onward distribution. This enables downstream measurement of consumption and, through the registry, the ability to "roll-up" total audience to the master content or brand.

Red Bee are responsible for the play-out of a large number of linear channels. UK broadcasters are mandating the provision of all content in a Digital Production Partnership AS-11 standard package that includes descriptive metadata and unique IDs. As channels shift their ingest process, Red Bee will be able to automatically match incoming media to its extensive library. US film and TV studios are testing their DPP packages including EIDRs over the summer in anticipation of the Oct 4th deadline.

Smart TV guides– we have a multi-year deal with Samsung to provide rich metadata services for Samsung's new range of Smart TVs across 18 European countries. We provide rich EPG content, including localised programme descriptions and images, to power content discovery for over 2,500 channels in 31 languages across those European countries including the UK, France, Germany, the Netherlands the Nordic region and Eastern Europe. The service makes TV content easily searchable on Samsung Smart TVs, and allows users to discover content through customised recommendations and voice navigation. Much of the popular content will already have EIDR ID's, where European content from smaller providers has not been registered Red Bee plan to create new registrations where possible for uniform presentation of EIDR IDs across the media delivery.

Red Bee also plan to provide a "service bureau" for content owners that do not have the necessary infrastructure or technical resources to register their content libraries and maintain the rich metadata sets necessary to provide the latest generation applications across multiple devices for platform operators. Content distribution deals often include the provision of rich metadata from the content owner which is then moderated and rewritten into the "tone of voice" required by a platform such as UPC.

Next steps

Further development will fully embed the EIDR API into the TeleView UI so it becomes the business-as-usual process to register premium content as we build metadata and reference other systems using Alternate IDs. This could include tapping into linear broadcast workflows to reuse IDs.

We will also continue to participate in "tiger team" web conference sessions that discuss and agree the development of best practice and new system requirements. As a member we can make sure that our particular European client requirements are taken account of in new developments.

OTHER EUROPEAN EXAMPLES

The British Film Institute is the worlds largest film and TV archive, they have started the process of matching and registering the entire BFI catalog with EIDR starting with the complete UK filmography by decade and by collection (Hitchcock, Bond, Kubrick, Hammer Horror, etc). Using the desktop XML tools this adds a BFI ID in EIDR as Alternate ID and the EIDR ID in the BFI record on the Collections Information Database. This process acts as an impetus to improve the core BFI metadata and make it more available to researchers and for public access. With the Alternate ID, EIDR provides a link back to the BFI record, and eventually BFI will include a link back to the EIDR registry from BFI web searches. The EIDR ID is becoming a part of existing BFI processes, for example new BFI-funded films or content digitised through the Lottery funded Unlocking Film Heritage program

British Audience Research Board - EIDR ID's are being considered as part of the content metadata in the new BARB Dovetail project that will combine the current "gold standard" TV panel for audience measurement with on-line players for the major broadcasters

StudioCanal - Technicolor are in the final stages of a trial project to register StudioCanal content, this integration could be extended into Canal+ broadcast operations

UK Intellectual Property Office -The recent report "Copyright Works: streamlining copyright for the digital age" stresses the importance of interoperable standard identifiers for media. It urges UK Broadcasters and content owners to consider DOI-based solutions, and makes note of EIDR-ISAN interoperability.

LATEST US ACTIVITY

Alternate ID Matching - As the US studios integrate the API into their standard internal workflow systems they are registering Alternate IDs in bulk, there are now more Alternate IDs in the registry than EIDR IDs with some records having over 20 different external IDs. Baseline uses EIDRs matched to its internal ID to automatically provide enhanced metadata sets to its studio and platform customers.

Digital Distribution – Content is increasing being consumed on a range of connected devices including powerful games consoles. Consumers use Xbox Live to find games and entertainment content ranging from ESPN3 to Comcast's Xfinity Video-on-Demand library. Warner Brothers and Microsoft built a new operation workflow streamlined using EIDR for the online retail distribution of entertainment assets. They contracted an external consultant to identify savings in:

- Asset Ingestion/Manual Quality Control : by using EIDR ID for avails and ordering, a saving of 270 hours/year per content partner
- Manual Communication: by using EIDR ID for ordering, delivery and confirmation, a saving of 375 hours/year per content partner
- Delivery Reporting/ Asset Queries by using EIDR ID for avails, ordering, a saving of 7 hours per year, plus fewer Customer Queries and delivery distractions in day

UltraViolet - Sony, Warner Bros, NBCU, and Fox are creating EIDR IDs for titles and individual edits derived from the titles to support digital distribution. EIDR IDs, along with associated metadata, are pushed from a content owner's internal title management system to a Content Distribution Network (CDN) provider. Assets are paired with the EIDR and pushed as a package into the UltraViolet workflow. The UltraViolet coordinator sets digital rights locker management and publishes to retail partner sites, where customers then buy rights to the movie or TV titles.

TV Everywhere and Anytime - Comcast, Charter, Cox and Shaw are distributing to multiple consumer devices – tablets, PCs, phones – in a trend known as TV Everywhere. The platforms need to track the VOD assets end-to-end as they traverse back office, distribution, and consumer/device environments. Studios registers content with EIDR, then delivers those assets, metadata and the associated EIDR to an on-demand delivery partner. That delivery partner in turn retrieves EIDR metadata directly from the registry, and provides EIDR IDs in the CableLabs VOD Asset Distribution Interface (ADI) specification feeds to the MSO. The cable provider brings the EIDR ID into its internal database and correlates asset metadata utilizing EIDR, linear vendor IDs and others. The MSO provides VOD results with EIDR to a usage measurement vendor, which produces VOD usage/purchase reports utilizing EIDR for better reporting and measurement results.

EIDR IN THE STANDARDS LANDSCAPE

ISO/DOI - EIDR is based on the Digital Object Identifier (DOI), ISO standard (ISO 26324)[4]. The DOI syntax is also an ANSI/NISO standard. Many industries use the DOI standard and infrastructure to create persistent, resolvable, interoperable identifiers; examples include CrossRef for scholarly publishing and DataCite for scientific datasets. ISBN numbers for books can also be represented in the DOI infrastructure.

DVB – EIDR is now included in the Companion Screens standard [6]

AlternateIDs - AlternateIDs may include, for example, CRID (RFC 4078), ISAN, ISRC, SMPTE-UMID, UPC, URI, as well as private or commercial ID systems such as Ad-ID, Baseline, IMdB, Muze, etc.

Mapping from Other Identifiers to EIDR - Other existing IDs can be directly mapped into EIDR. EIDR has published a document that lists all of the supported Alternate ID types, as well as suggested mappings from

existing IDs to EIDR, called EIDR Registrations from Other Systems at the EIDR Web site. Detailed mappings are described for the following standards: .

- International Standard Audiovisual Number (ISAN) - ISAN is widely used in rights management and collection systems. A complete mapping of an ISAN registration to an EIDR registration is possible today, and a joint testing effort is underway. Thus, users of either system are able to operate in various domains as needed, using the identifier of their choice for any particular application. ISAN and EIDR are continuing to work on best practices, guidelines, and other technical documentation to further facilitate the interoperability of the two registries.
- EN 15907 and EN 15744 - These standards are under the auspices of the European Committee for Standardization CEN/TC 372 and filmstandards.org. Best practices and mappings are available for EN 15907 and EN 15744 root objects. EIDR is also working with film archives to extend interoperability with these standards to a more granular level of detail.
- EBUCore is a common core set of descriptive and technical metadata that describe media resources (audio, video, audiovisual, still images, subtitling, etc.). EBUCore is provided for free use by the European Broadcasting Union. A mapping of EBUCore for base records to EIDR root objects is provided. EIDR is working with the EBU and others in the definition of SMPTE Core.
- SMPTE – RP 2079:2013 describes the standard of encoding of EIDR and DOI IDs for use in other SMPTE specs, e.g. MXF, AS-03, and AS-11. RP 2021-5:2013 specifies the use of EIDR and Ad-iD in SMPTE BXF and ATSC PMCP.
- CableLabs Metadata standard for the distribution of video on demand assets [2], EIDR is one of a half dozen program identifiers that can be present in an SCTE-35 segmentation descriptor, the standard for digital ad insertion in cable distribution.
- Ad-ID - The advertising Digital Identification LLC (Ad-ID™) advertising asset identifier and metadata is a fully compatible Alternate ID within the EIDR structure. The interoperability of EIDR and Ad-ID has been proven out in CableLabs industry testing events as a means of targeting advertisements to relevant content assets. The CIMM/TAXI project in the US showed the value f being able to link ads and programmes together using Ad-iD and EIDR.
- IANA/IETF – the URN form of an EIDR ID has been approved by IANA and publication of the format as an IETF RFC is underway.
- Linked Heritage, European Union -Linked Heritage is producing a mapping from EIDR to LIDO as part of its charter to coordinate standards and technologies to increase the quantity and quality of metadata available to Europeana. This in turn will make it possible to expose descriptive metadata for culturally relevant products alongside links to purchase them from retailers.

CONCLUSIONS

EIDR usage has grown rapidly in the US and is being applied in the UK and across Europe. While the concept of a Universal Content ID is intuitive and obvious, its implementation requires changes in existing workflows and systems. The use of the web registry and mapping to alternate ID's significantly reduces the set-up and ongoing costs of integration. Business approval for those changes requires real-world worked examples of the commercial benefits. When the critical mass of content has been mapped with EIDR IDs within a domain a range of new applications can be tackled that rely on its "metadata dial-tone" as an always available service.

REFERENCES

1. EIDR Reference documents http://eidr.org/technology/
2. Streamlined Content Metadata Integration and Management Using Entertainment ID Registry (EIDR) A Technical Paper prepared for the Society of Cable Telecommunications Engineers By David Agranoff Senior Architect CableLabs http://www.eidr.org/documents/Expo2012_Technical-PaperEIDR.pdf
3. MovieLabs Common Metadata 2.1 (md: prefix) as found at http://movielabs.com/schema — generically (documentation at http://movielabs.com/md/md/)
4. DOI Metadata Kernel (doi: prefix) as found at http://dx.doi.org/10.1000/276
5. DOI version of ISO 3166-1 alpha 2 code as found at http://dx.doi.org/10.1000/279
6. TM-CSS0130r22_Draft_TS_102_CSS_Companion-Screens-and-Streams