

TF-netcast deliverable G: Report on video-on-demand metadata and portals

Harri K. Salminen, CSC / Funet-TV

<http://www.csc.fi/staff/harri.salminen/>

Purpose of this document

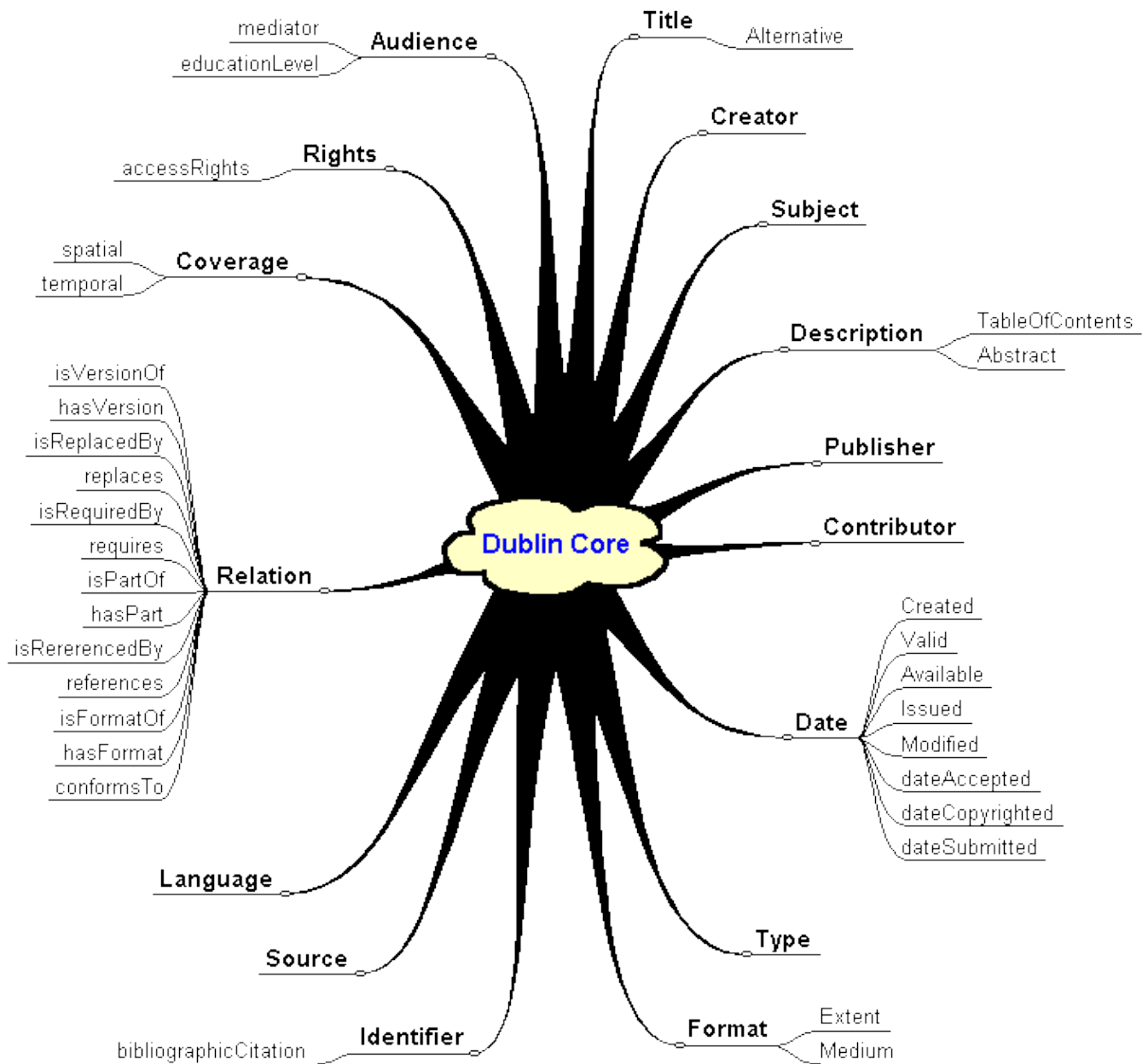
This document reviews several metadata models that are currently used for describing video on demand assets in the academic community and then explores how this metadata could be exchanged for use in a portal.

Review of existing metadata models

Video on demand assets have been described at varying details with several different metadata models that range from generic models suited for any kind of objects to models specifically designed to describe minute details of video assets but not well suited for anything else. In the tf-netcast [Report on Streaming Video Survey](#) 30% of the respondents said that they use metadata to describe their assets and the most popular one was QualifiedDublin Core. The second most popular was the IMS, other standards got support from one or two respondents.

Dublin Core (DC)

[Dublin Core](#) metadata initiative has developed a recommendation for a least common denominator among various metadata standards. It's core is called the Dublin Core Element Set that consists of 15 fairly generic elements that are all optional and repeatable. The idea is that these semantic elements could be used to roughly describe almost any kind of document for the benefit of building union catalogs or other metadata exchange. Though originally developed by the library community with orientation to printed documents it has then been adopted as a common metadata format for web pages and many other electronic assets including video. Since all elements are optional, one can usually always find at least something semantically equivalent elements like Title or Identifier in almost every other metadata description that could meaningfully be mapped to Dublin Core.



The original fairly purist 15 element unqualified Dublin Core hasn't been detailed enough for many needs so it has been refined by adding terms and qualifiers that add more semantic detail to the elements. This is often called as Qualified Dublin Core. Currently the standardized qualifiers and one new element, audience, have been integrated with the Dublin Core Element Set in the recommendation called DCMI Metadata Terms. This document also specifies names for standard encoding schemes for some of the element values as well as a type vocabulary. One should note that in September 2003 it was finally decided to split the type Image to Still Image and Moving Image which undoubtedly will help us to restrict searches and harvesting to only video assets.

Since Dublin Core has wide acceptance and good basic semantics it has been supported at some level by many of the current metadata databases used for VoD assets in the academic community. Some may support only the unqualified Dublin Core but many have used qualifiers or extensions to support local needs.

Videnet has developed a Dublin Core Application Profile for Digital Video that refines the DC specification with many qualifiers and subelements. It also adds a new root level element hierarchy for metametadata, that is metadata about the metadata record itself, not about the object being described. It's useful for the management and distribution of the metadata descriptions for example with the Open Archives Initiative protocols. The Videnet profile document has many useful comments on how Dublin Core can be used for describing Digital Video assets. They have a MS Access 2000 metadata record specification but they don't seem to provide an XML Schema, only examples in both XML and RDF syntaxes that can be used in designing your own XML schema or mappings. The Videnet profile is thus a good starting point for those who want to use Dublin Core for video specific information.

Surfnet Video Portal metadata model is basically qualified Dublin Core with dutch translation and some fields made mandatory. This model has also been adopted for use in Denmark.

The metadata of DSPACE at MIT uses dublin core with some custom refinements they needed. These seem to be mainly for compatibility with the local library and document management needs and not for Digital Video like the Videnet extensions. They have hoped to add support for IMS/SCORM during 2004 and research possibilities for arbitrary schemas by using RDF.

CSC's Funet-TV has developed its own XML Schema based on the Videnet's Dublin Core for Digital Video XML descriptions with additional refinements and extensions to support full LOM semantics, OAI, some MPEG-7 concepts for timecodes, persons, groups and organizations, VCARD descriptions in XML and some local refinements like keyframes, icons, rich descriptions and cast in the DC's description element. To support multiple alternative formats or locations for an asset they are referred to by extending the relation/hasformat to a hierarchy supporting necessary technical details. This metadata description is used currently only by the Funet-TV's own open source based mediar-project. Mappings to a simple unqualified Dublin Core or standard qualified Dublin Core are fairly easy to implement since OAI compatibility was taken into account in the design which includes the metametadata elements.

Learning Object Metadata (LOM)

IEEE has developed a Learning Object Metadata or LOM standard to describe entities that can be used or referenced during technology supported learning. It has also been adopted with some additions and modifications by the IMS Global Learning Consortium, Inc as Learning Resource Metadata which specifies one possible XML binding for LOM metadata. IMS metadata specifications have in turn been adopted by the Advanced Distributed Learning Initiative as part of the Sharable Content Object Reference model (SCORM) to define Web-based learning "Content Aggregation Model" and "Run-Time environment" for learning objects. In Europe the Ariadne Foundation that resulted from the EU projects for "Telematics for the education and training" co-operate currently with both IMS and the ADL Initiative in the development and use of LOM. The Finnish broadcasting corporation has a pilot project in association with the Finnish Information Society Development Center (TIEKE) to test the Finnish translation of LOM which has been experimentally mapped to the Funet-TV's extended DC metadata.

LOM or its derivatives can be used for describing educational video assets although it is not specifically tailored for them. It's not meant for scientific assets or generic library databases either thus it's mainly used by user communities specializing in electronic learning objects. However, the IEEE LOM working group and the DCMI have signed a Memorandum of Understanding that outlines the commitment to develop interoperable metadata. Already a mapping to unqualified dublin core does exist and majority of the other LOM elements can be mapped to qualified or at least extended dublin core. This means that LOM, IMS and SCORM conformant databases could be programmed to export a dublin core metadata description that's usually good enough for building common portals or union catalogs. Conversely, the Dublin Core descriptions of a video

could be converted to LOM and the missing elements could be filled locally or left empty.

MPEG-7

The ISO/MPEG-7 standard finalized in 2002 after many years of work by large number of international experts has been specifically developed for the search and retrieval of multimedia data in minute detail. It has also been called as "Multimedia Content Description Interface" and it's expected to be adopted especially by the digital television and streaming industry in the future. In addition to the usually manually entered descriptions of titles, keywords, persons, dates, rights etc. MPEG-7 supports automatically generated metadata about for example camera movements or timecodes. What sets the standard apart from other metadata standards described in this report is the ability to support audio and visual descriptors. Audio descriptors can be used for example analyze and describe the sounds, speech or music in the multimedia asset and perform for example queries by humming. Video descriptors can be used for example for face recognition or queries by example.

The standard is quite complex with 1352 pages. MPEG-7 is described using the Description Definition Language (DDL) which is basically XML Schema description language with extensions to support array, matrix and some temporal data types. The metadata is XML conformant even though full schema conformance verification will require a special parser. As XML it can be exchanged and processed by many common tools. MPEG-7 can also be represented in a binary format (BiM) which is especially important for devices with limited processing resources and bandwidth. BiM might for example be used along with other binary MPEG formats in embedded applications like mobile phones or digital televisions. As with other ISO standards, they are not freely available like most Internet standards but sold commercially (1469 CHF for the whole set) which may limit at least low budget open source code production for it. To understand it's possibilities there's a 79 page [MPEG-7 overview](#) available on the web. Some of the standard's developers have also written a 352 page book called "Introduction to MPEG-7: Multimedia Content Description interface" (ISBN 0471486787) that helps in understanding the MPEG-7 standard. To fully implement the standard may require fairly substantial amount of resources which could slow it's adoption in small scale academic projects in the near future. So we need to wait for the large companies to produce MPEG-7 aware products or start by implementing some small part of the standard as has been done for example in the [MIC Union catalog MPEG-7 mapping](#). A very small subset can also be mapped to unqualified dublin core thus enabling wide interoperability.

MARC 21

The MACHine-Readable Cataloging or MARC development was started about thirty years ago to enable the US Library of Congress to exchange bibliographic information with other institutions. The MARC 21 version came out in late 1990's and it's been widely used in scientific libraries all over the world. To store metadata values it uses fields identified with three digit codes and possibly additional subcodes. The bibliographic metadata portion of the standard has about 4000 fields and subcodes intended for precise description of bibliographic information of mainly physical objects. For video there's support for motion picture reels or video tapes but apparently not specifically for streaming video formats, thus it doesn't in practice compete with MPEG-7 in that arena. It's an extensible standard so perhaps more support for virtual objects like streaming video will appear someday.

In the [CERN Document Server](#) streaming videos have been described with MARC like any other bibliographic object with some special coding to create the necessary URLs for example. There does also exist crosswalks to both unqualified and qualified Dublin Core from MARC which probably would need modification for VoD portal interconnection. Especially when exporting to DC one should take care that there's an URI that can be used to locate the online objects even though it might not be necessary inside the

local database. Since it's a fairly complex and library oriented standard it's best used in association with a scientific library and professional librarians. Concise descriptions of the MARC standard and it's new XML binding are available from the [Library of Congress – Network Development and MARC Standards Office](#).

Metadata granularity

One often neglected issue that becomes apparent with video clips is the need for different levels of granularity on the metadata descriptions. E.g. there could be a whole conference or course series, that consists of many different presentations which in turn may contain several different multimedia assets that may even have a storyboard or synchronized lecture notes to facilitate finding a specific point of interest in a long stream of video. It becomes more important when the collections grow so at least some levels of granularity should be supported from the beginning.

In the standard Dublin Core there's no notion of granularity but one could use the relation and possibly subject fields to support at least some sort of collections. In the LOM metadata model there's support for the granularity attribute which can have several integer values depending on the level of granularity:

- 1 Atomic level: Videoclips or files
- 2 Collection of level 1 objects: lecture, movie
- 3 Collection of level 2 objects: Course or series
- 4 Collection of level 3 or 4 objects

In the Funet–TV metadata model, which encompasses LOM, it has been slightly extended to allow also subatomic levels if necessary. MPEG–7 of course excels in the representation of internal structure of a video. MARC 21, being a library standard, is often used to represent aggregation in the form of a series of publications, separate collections or as classification hierarchies of which UDC and DDC are the most well known ones. Both UDC and DDC are however copyrighted and not freely usable in internet portals without special agreements which severely limits their use in open VoD portals.

If the metadata repository supports the notion of hierarchy to represent a collection of metadata documents I'd recommend a naming convention based on reverse domain names, if applicable, of organizations responsible for maintaining the original information. This would mean creating globally unique prefix paths before locally defined hierarchical levels for such information. E.g. fi/csc/seminars/xyz/abc.xml or net.terena.task–forces.tf–netcast.demos.video1. In this way naming clashes between different metadata origins are easily avoided and replication can often be done very easily and efficiently. This kind of administrative hierarchy could also be used for aggregating assets by creating metadata documents that describe each level of hierarchy. E.g. the Funet–TV medar database uses a file called index.xml to contain metadata that is common to all other metadata descriptions in the same subdirectory, which could be an introduction to a course or a conference for example.

Metadata identifiers

Title, publisher, creator, URIs to the video instances etc. taken in combination describe an object uniquely, but they are not necessarily persistent or easily related to the dataset to be replicated which makes automatic processing and duplicate detection very hard in a distributed heterogeneous metadata environment as well as creates problems locating the object referenced in a web page or a printed document after the object's location has changed. To solve this issue globally unique relatively stable identifiers should be provided for each unique object. It might be enough to have a unique identifier for each metadata description as long as it's not directly tied to a particular database or protocol that might change over the years.

Books usually have a ISBN number and a similar URN scheme has been promoted for many years by libraries for electronic publications and other stable electronic documents (see [RFC 1737](#)). One URN implementation well suited for automation in the metadata creation process is the [Digital Object Identifier \(DOI\)](#) which is gaining wide acceptance globally. To resolve the DOIs currently the CNRI's open source [handle system](#) is being used and you need to register an organizational prefix. It has been designed as a system independent long term identifier that's easy to deploy and resolve and already over ten million of them are in use especially in scientific community.

The Open Archives Initiative recommends it's own [OAI identifier](#) with it's own registration scheme but other unique URI scheme's could be used instead if seen necessary. OAI identifiers are actually designed to be the unique id of a particular OAI metadata record regardless of it's representation and location and not as an system independent unique identifier of the object being described. Also they use a Domain name as the organizational prefix which doesn't require registration but on the other hand may not live as long as the DOI prefixes since they can time out. Thus I'd recommend using DOI instead or in addition to the [OAI Identifier format specified in their implementation guideline](#).

The main thing to note is that you first need to select one or more URN schemes within which you to register your system with a unique administrative domain id. Then you need a stable unique identifier within your system which could be a random code or even filename and path, if it's kept stable. The combination of the scheme, domain and local identifiers can then be used to form a URN according to a specific syntax related to a scheme. So it's technically fairly easy to support multiple scheme's provided that you can administratively keep their meaning same and stable.

Examples of possible identifiers

oai:tv.funet.fi:fi/csc/seminars/xyz/abc.xml
oai:terena.net:net.terena.task–forces.tf–netcast.demos.video1
oai:csc.fi/1234567890987654321

doi:10.1234567/fi/csc/csc/seminars/xyz/abc.xml
doi:10.12345678/net.terena.task–forces.tf–netcast.demos.video1
doi:10.123456789/1234567890987654321

Methods for metadata exchange

Users don't usually don't have time to go to look at many different places in case they happen to have something they are interested in. The large web search engines have solved this by trying to index everything for everyone automatically which requires huge resources and usually lacks support for local metadata or preferences. In the Terena community different organizations may have quite different local needs and solutions for organizing metadata. Thus to create single search portals we need to interconnect those local metadata repositories in one way or other. For interconnecting metadata repositories there's two basic approaches: query forwarding and replication of metadata.

Distributed querying

In the query forwarding method a portal just forwards the search or browsing request to remote repositories and then presents the combined results to the user. Simplest implementations just divide the web page to multiple frames with a different web interface of a remote search engine in them but such an approach does not scale well for more than few databases. More advanced approaches use some sort of common protocol to do the queries in background and present the results in a single result list to the user.

The [Z39.50](#) (ISO 23950) standard specifies a client/server-based protocol for searching and retrieving information from remote databases. It has been used for many years by the library community but possibly due to its complexity it has not been widely adopted for portal building in other communities. Thus there's now a "[Z39.50-International: Next Generation](#)" or ZING initiative to make it more widely used for example by developing a new interfaces and a Common Query Language (CQL) by using XML technologies and SOAP protocol instead of the original ASN.1 coded OSI inspired binary protocols. These should make it easier to query large and complex library databases from portals. Especially should be noted that the Search/Retrieve Web Service protocol should support common XML schema for Dublin Core thus making interoperability with less complex systems easier. Several other database implementations do however implement simpler custom query protocols using XML or some other representation which of course leads to the need of custom integration or conversions.

The problem with all these distributed query approaches is the need to query in real time multiple remote databases with reasonable success and response times. Although many database queries parallelize quite well in large web search engines like google it's harder to do fast enough in a heterogeneous environment without common management.

Metadata replication

To avoid the problems associated with distributed querying library community has developed the concept of union catalogs that contain only essential information in a common format for finding the right database with detailed information. This union catalog is often based for example on MARC or Dublin Core metadata and is on a centrally maintained database. E.g. the Finnish scientific libraries use this approach to collect MARC 21 records to a central union database.

In the US the [The Moving Images Collections \(MIC\) Portal](#) has a goal to create a union catalog as window to the world's moving image collections. It has a custom core metadata schema that supports importing of relevant common elements from the Dublin Core, MPEG-7, Videnet DC and MARC 21 standards for common retrieval. This project was initiated by the [Association of Moving Image Archivists](#) by a [National Science Foundation \(NSF\)](#) grant and the goal is to host the finished database at the Library of Congress.

In the metadata replication method a local copy of the contents of a remote repository is made at the extent that's deemed to be of interest to the local portal maintainers. If the remote repository is a database that can return complete search results in a machine readable format like XML using a known schema, then such queries could be made every night and stored in a local repository. One approach available with several search engines is to formulate the query as a URL that returns all search results in some suitable format which could be XML or even HTML. In case the result set is very large, it may create technical problems, but otherwise it's usually quite simple to implement in existing web based search engines. A more sophisticated approach might use a XML and HTTP based transactions to retrieve the information that should be replicated. Web Services use the SOAP protocol but other protocols are also around.

Open Archives Initiative

The [Open Archives Initiative](#) (OAI) has defined a protocol especially for exporting XML-based metadata from open repositories to harvesters which is supported by several academic metadata providers including some video on demand providers. OAI supports retrieving of metadata records for the whole repository or a subset. The subset can be defined either by a time range or custom predefined set identifier which might map to some database query or directory tree in a filesystem based repository. OAI requires that all metadata is at least available as simple unqualified Dublin Core. It does however allow the use of any other XML schema for optional representations. It is much simpler to implement than Z39.50 and there's already many open source

implementations available. OAI is supported already by for example DSPACE, MIC and CERN CDS.

The unqualified dublin core records transferred with OAI could contain an URL pointing to the complete metadata record in the original database thus making it easy to create local custom union catalogs in each portal without necessarily a need for a central union catalog like MIC. This two click approach would also make sure that the user always knows where the video is coming from and provide all the necessary information to access the asset which may be available in different formats from different media servers.

File replication

If the remote repository's contents can be represented as files, then there's many ways for replicating them efficiently using already well established protocols and open source tools. If the repository is represented as a set of interlinked HTML pages or there's a directory index or a single search result page for the files then it could be replicated automatically. The problem with replicating metadata as HTML pages is that you'd probably need to parse the pages to extract the locally needed information in a proper format (e.g. valid XML files) and even a single change in the style or representation of the web pages or links might break the parsing. HTML does support simple metadata by the use of tags in the section but it has been misused so much that the big web search engines don't use it anymore. Thus I'd recommend HTML replication mainly for information that should remain as HTML and not for metadata that needs to be parsed unless there's no better alternative. This often applies to different databases with web only interfaces that you may want to mirror as files for inclusion in your local portal application.

File replication tools

File replication has been done in large scale for distributing freely available software between anonymous ftp archives like ftp.funet.fi for many years. One of the oldest protocols in the Internet for file replication is FTP. Although originally designed as an interactive protocol for human users it can be automated with special mirroring programs that simulate a user and replicate all changed files in a specific remote directory hierarchy to a local one with original date information.

Most ftp servers and batch clients are designed as freeware for Linux and Unix operating systems, which includes Mac OS X. If you don't already have a preferred ftp server you may want to consider [Pure-FTPd](#) which is secure, reliable and fairly easy to use. It supports for example privilege separation, chrooted virtual users, ssl, secure uploads, bandwidth limiter, IPv6 etc. and it should even work under windows if you really need that.

For mirroring many custom software solutions have been developed over the years which may not be the best choice for beginners. However many command line unix ftp clients like [ncftpget](#) and even browsers like [lynx](#) do support batch operation. However, for mirroring large directory hierarchies you should use software that supports copying only those files that are changed like [lftp](#) which can also mirror with ftps, http, https, ipv6 etc. You can find it and others via [Linux.org application directory](#). Other possible freeware alternatives might be web mirroring tools that also support ftp like [wget](#) or [HTTrack](#).

[Rsync](#) is a freeware client and server which implements a protocol that's designed for efficient and reliable directory hierarchy replication. For example it can copy only changed portions of large file if necessary and it is being used to mirror sites with hundreds of thousands of files and megabytes. It works best with Linux / Unix and requires only a single tcp port to be opened. It can be run through ssh if a secure connection is required or even used locally to mirror files from a locally mounted filesystem.

There's also several other methods for file replication like various network filesystems and P2P technologies which may fill in some special needs. In any case the end result is usually the same, you have a local copy of a

remote directory hierarchy. One should note that the file replication methods can also be used to replicate the assets itself to another media–server for a distributed service. In this case authentication and possibly caching methods may need to be implemented to enforce copyrights or limit resource usage.

Conclusions

Video on demand services are being set up all over the internet but there's currently no easy way to query or browse their combined contents in a consistent fashion. Automatic web–crawling robots of commercial search engines like Google or Singingfish may find a significant number of the resources if they are suitably exposed on simple web pages and widely linked to from other sites. Many popular search engines don't even try to use metadata on web pages since it has been misused a lot which limits the usability of searches making it complicated to just focus on videos from universities related to a particular scientific subject.

There's already some interest in exchanging metadata descriptions for video assets between various organizations both in the Terena and Internet 2 communities. I see several possibilities to solve the problem. The seemingly easiest way is to delegate the problem to the libraries which already have databases and probably use complex metadata like MARC 21 to describe all kinds objects. However, in many organizations, they might not yet have resources or interest to catalogue every video clip that the users may wish get published. Some organizations don't even have a professional librarian or central library database.

This has lead to the development of local video on demand portals with different metadata description schemes. The only common denominator seems to be some sort of possibility to support at least unqualified Dublin Core. Many support similar refinements but for the widest interoperability they could be mapped to the unqualified forms in a consistent fashion. E.g. given names and surnames probably should be in an order that facilitates alphabetic indexing. When unqualified Dublin Core is used and the site actually supports richer metadata format or multiple different streaming formats then the identifier elements should contain a URI pointing to a standard conforming web page with full details. In theory it should be enough to have a URN, preferably DOI, as the only identifier but in practice one should also include beside it a simple URI with http method that most browsers can use directly. However since there does exist several systems that already support quite similar richer metadata format be it qualified or extended Dublin Core, LOM/IMS/SCORM, MARC 21 or MPEG–7 one should provide alternatives to exchange the richer metadata formats between those systems with descriptions of the particular schemas used. The metadata records should still use same DOI as in the case of unqualified Dublin Core.

To achieve widest interoperability among various databases at least unqualified Dublin Core with URIs that be resolved by a standard web browser should always be supported. For a richer description qualified dublin core or extended Dublin Core that refines the ambiguities allowed by the standard dublin core is a popular alternative. Good starting points for those wishing to extend Dublin Core for video might be for example the Videnet and Funet–TV metadata schemes. In any case, it's recommended that the local schema or metadata usage guidelines are made publicly available.

In the longer term the MPEG–7 standard should gain more popularity and you can already use some subset of it as a standard alternative for extending Dublin Core as has been demonstrated by MIC. The MPEG–7 related product development and adoption should of course be at least followed regularly but it may still take some time before it solves our video metadata interoperability problems. And you'll still probably need a good Dublin Core, MARC or LOM mapping for interoperability with union catalogs and other databases that describe also other than multimedia objects.

For exchanging the metadata the well supported OAI protocol is recommended. It can transfer in addition to unqualified DC any rich metadata descriptions with a XML schema. It's identify request supports description

TERENA TF-netcast deliverable G: VOD-metadata and portals

of the repository itself as well as pointing to friends with OAI repositories thus making it easier to form a network of repositories to harvest from. OAI would make it also technically easy to collaborate with the US metadata exchange efforts concentrated around the MIC project.

File replication is even simpler alternative to OAI if the whole repository's contents can be exported as XML files or you need to exchange local copies of other files like HTML pages, presentation materials or even the video files itself. For replication rsync is probably the most efficient protocol but either FTP or http replication might be easier to arrange and still fit the needs.

To support the DOI resolution the CNRI Handle system should be installed and administrative policies implemented locally to maintain stable digital identifiers. We can start without them by using OAI identifiers or just some local URIs but sooner or later a standard URN implementation should be adopted.

When the number of available repositories grows we should develop a specific metadata schema for detailed description of the available metadata and asset repositories including what are the protocols supported, limitations of use, available metadata schemas, vocabularies and taxonomies used, communities or subjects covered, possible proxies and mirrors etc. This could then be used to automatically update information on the availability of remote repositories for possible replication to the local repositories be it a simple server with static files or a full blown library database. Authenticity of such metadata should either be provided by using a trusted hierarchy of servers, possibly including a central union catalog, as intermediaries or public key authentication methods with a web of trust.

If you don't already have a suitable portal system available, then you might want to take a look at the open source DSPACE software that already supports OAI with unqualified Dublin Core as well as DOI. You could also ask from the people behind other portals you like what software they use and whether it might be freely or commercially available. If nothing else, at least try to use consistently some metadata format even to just create web pages by using some simple scripts. Almost any kind of metadata will save lot of effort as soon as a need for automated processing of the data arrives be it local web page redesign or exchange of metadata with others.