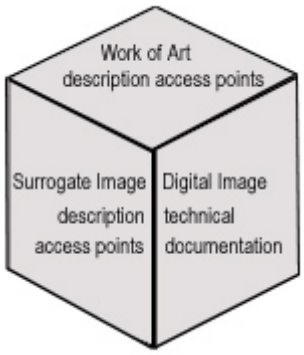
## Selecting a Metadata Schema

There are many metadata schemas available, geared to different communities and to different needs. Metadata schemas are defined ways of structuring metadata elements; in other words, they are used to structure information or data. The idea behind the development of metadata schemas is to promote consistency and uniformity of data so that it can be easily aggregated, moved, shared, and ultimately used as a resource discovery tool or for other purposes. Participation in resource-sharing or collaborative initiatives often requires the adoption of particular metadata schemas.

Fig 14. Model of metadata required to document a work and its surrogates




One of the complicating factors in documenting hybrid collections is that there are likely to be many versions of one object, and their relationships to each other must be recorded. For instance, a collection might include an original painting, a photochemical surrogate of the painting (such as a 4-by-5-inch transparency), and several digital surrogates derived from the transparency: the archival master, the derivative master, and one or more access files. Some metadata schemas provide 1:1 documentation of each version, while others are hierarchical, designed to provide information on parent, child, and peer relationships without requiring that each instance be separately and fully catalogued. In fact, many schemas that were originally intended to provide 1:1 documentation are now being extended to accommodate this new reality, but the exact manner in which this is done varies and is again the subject of current research.[3] Broadly speaking, developers and users of documentation for image collections distinguish between information that refers to the original work, information that refers to analog or digital representations of that work, and information that refers to the technical characteristics of a digital image (which may have been captured from the original or from a photographic representation or derived from an existing digital image file) (fig. 14). They also distinguish between schemas designed to describe individual items and collections as a whole. Various schemas may be used together to provide complete documentation of an item or collection.

**MARC** (Machine-Readable Cataloguing) is a venerable metadata standard long used for creating bibliographic records. It supports the Anglo-American Cataloguing Rules (AACR2) and allows the exchange of cataloguing information within the library community. MARC has been enhanced over the years to, for instance, accommodate the cataloguing of nonbibliographic material, represent authority information, and include elements to describe electronic resources. The Library of Congress and the MARC Standards Office have developed an XML schema, the Metadata Object Description Schema (**MODS**), designed to both transmit selected data from existing MARC 21 records (so-called because they result from the harmonization of the Canadian and U.S. MARC formats in readiness for the twenty-first century) and enable the creation of original resource description records.

**EAD** (Encoded Archival Description) is a set of rules for creating finding aids for archival collections that specify the intellectual and physical arrangement of an intact or cohesive collection as a whole. EAD finding aids may be

linked to item-level records that exploit **Dublin Core** or some other schema. EAD uses **SGML** (Standard Generalized Markup Language) to define its logical structure (see *Metadata Format*).

Dublin Core, developed as a core set of semantic elements for categorizing Web-based resources for easier search and retrieval, has become popular in the museum and education communities. The schema is deliberately simple, consisting of fifteen optional, repeatable data elements, designed to coexist with, and map to, other semantically and functionally richer metadata standards. Dublin Core's simplicity makes it an excellent medium of exchange, and thus a basis for interoperability. The metadata harvesting protocol of the Open Archives Initiative (**OAI**), known as OAI-PMH, which provides a mechanism for harvesting or gathering XML-formatted metadata from diverse repositories, mandates Dublin Core as its common metadata format. The Dublin Core Metadata Initiative (**DCMI**) is also developing administrative and collection-level metadata element sets, and user communities have developed element qualifiers relevant to their own fields, which both enrich and complicate the standard.

**CDWA** (*Categories for the Description of Works of Art*) provides a framework for describing and accessing information about artworks and their visual surrogates. It identifies vocabulary resources and descriptive practices intended to make information held in diverse systems both more compatible and more accessible. The Visual Resources Association (VRA) Data Standards Committee expanded upon certain portions of the CDWA to formulate the VRA Core Categories, which are specifically designed to describe the attributes of not only original works but also surrogates in considerable detail. This makes the **VRA Core Categories** particularly useful in documenting digital image collections.

The Research Libraries Group's (**RLG**) **Preservation Metadata Elements** are intended to set out the minimum information needed to manage and maintain digital ?les over the long term and, unlike the schemas described above, capture technical, rather than descriptive, information. This element set may be combined with any descriptive element set to describe an image file.

The NISO *Data Dictionary: Technical Metadata for Digital Still Images* provides an exhaustive list of technical data elements relevant to all aspects of digital image management: preservation, production, display, use, and processing. In contrast to the RLG preservation elements, which can be applied broadly to many types of digital files, the *Data Dictionary* focuses only on digital still images. The Library of Congress and NISO have developed an XML schema known as NISO Metadata for Images in XML, or NISO **MIX**, based upon the dictionary.

**MPEG-7**, or Multimedia Content Description Interface, is an XML-based standard developed by the Motion Picture Experts Group (**MPEG**) to describe multimedia and audiovisual works and is likely to grow in importance over the next few years. It supports textual indexing (e.g., the use of controlled vocabularies for data elements such as subjects and genres) and nontextual or automatic indexing, such as shape recognition and color histogram searching. It also supports hierarchical or sequential description.

These are only a few of the metadata schemas available. Other metadata standards that might be of interest to the cultural heritage community include the *International Guidelines for Museum Object Information: The **CIDOC Information Categories***, developed by the International Committee for Documentation (**CIDOC**) of the International Council of Museums (ICOM). The International Council of African Museums (**AFRICOM**), originally a program of ICOM, has developed a data standard initially designed to promote the standardization of collection inventories in Africa, described in the *Handbook of Standards: Documenting African Collections*.[4] **Object ID** sets out the minimum information needed to protect or recover an object from theft and illicit traffic. **SPECTRUM**, developed by the UK-based mda (formerly the Museum Documentation Association), is comprised of a broad range of data elements associated with transactions for museum objects. CIMI (the Consortium for the Interchange of Museum Information) and the mda have developed an XML schema based on the SPECTRUM elements.

Broader systems are being developed that are designed to join the seeming morass of cultural heritage documentation, including the various metadata schemas, into a coherent whole, though many remain untested as of this writing. The CIDOC object-oriented Conceptual Reference Model (**CRM**), which took the CIDOC Information Categories as its starting point, is designed to mediate between disparate information sets by describing in a prescribed, extensible "semantic language" their explicit and implicit concepts and relations, thus promoting semantic interoperability. It focuses primarily on museum objects rather than digital surrogates but can be extended to allow for detailed description of such surrogates, for instance, by being combined with another metadata schema such as MPEG-7. The Metadata Encoding and Transmission Standard (**METS**) is a flexible XML encoding format for digital library objects that may be used within the Open Archival Information System (**OAIS**) reference model. METS provides a metadata "**wrapper**" that can hold together descriptive, administrative, and structural metadata and document the links between them, as well as capturing multipart

file groups and behaviors. Its generalized framework offers a syntax for the transfer of digital objects between repositories, while OAIS provides a common conceptual framework in which to understand the archiving of digital assets. (See *Long-Term Management and Preservation*.)

**Metadata Format**

While there are many ways to format the information captured by metadata schemas (see *Metadata*), the use of XML is becoming increasingly dominant. Like the more familiar **HTML** (Hypertext Markup Language), XML is a markup language derived from SGML. However, while HTML is designed to merely present data (as of this writing, it is used to format most Web pages), XML is intended to describe data: it is a hardware- and software-independent format that allows information to be structured, stored, and exchanged between what might otherwise be incompatible systems. XML gives users the ability to generate uniformly structured metadata packages that can be used for sharing; migrating; publishing to the Web; or archiving-such as for **ingest**, storage, and delivery in open archive environments. Even if metadata is not originally encoded in XML, it is likely that at some future point the need will arise to migrate or export it to this format, in order to maximize interoperability (many software manufacturers now support converting data into XML).

SGML is a complex standard that provides a set of rules for specifying a document markup language or tag set defining which coded tags and attributes may be used to describe a document's content. Such specifications are given in **DTD**s (Document Type Definitions); HTML is in fact a particular DTD, which Web browsers are designed to compile. HTML's tags are predetermined and largely limited to specifying format and display. By contrast, XML is a simplified subset of SGML. XML is "extensible" and flexible because its tags are unlimited, and thus anyone can invent and share a set of coded tags for a particular purpose, provided they follow XML rules.

XML and related tools can make semantic elements (such as "author" or "title") within documents **machine-readable** (i.e., allow documents to be parsed, effectively "understood," and acted upon by computer programs). XML documents are therefore potentially much more powerful than simple text documents. They are self-defining because they can refer to any DTD or **XSD** (XML Schema Definition), a more recently developed standard, to describe their data. DTDs or XML Schemas can define conditions such as whether a particular element is required, repeatable, or optional or express hierarchical complexity, such as parent, child, or group relationships. As compared to DTDs, XML Schemas facilitate information exchange between databases because they provide more sophisticated ways of structuring content and declaring data types. XML "namespaces" provide a means of distinguishing between duplicate element type and attribute names-derived, for instance, from different DTDs-and allow them to be used in the same document. Style sheets written in **XSL** (Extensible Stylesheet Language) dictate which XML data elements to show and how to display them. Extensible Stylesheet Language Transformations (**XSLT**) provides a standard way to reorganize the data represented in one XML document into a new XML document with a different structure or into another format altogether.

While XML is much more structured than HTML, the two are complementary; an XML Schema or DTD defines metadata elements that can be formatted for display using HTML or other formatting languages. **XHTML** (Extensible Hypertext Markup Language) is a reformulation of HTML as an application of XML designed to express Web pages and may be extended to include new elements.

XML could become even more powerful if routinely combined with RDF, which provides a model for describing Web resources that promotes a consistent representation of semantics. RDF effectively functions as a crosswalk between diverse metadata schemas and could allow structured metadata developed in different contexts to be exchanged and reused without loss of meaning. RDF provides a foundation for the machine processing of Web resources because it promises metadata interoperability across different resource-description communities and different applications. XML and RDF are both key components of the **Semantic Web**, the intelligent evolution of the World Wide Web proposed by its inventor, Tim Berners-Lee, but as yet not realized. [5]