Identifier Management and Resolution: conforming the IEEE Standard for Learning Object Metadata

Roberto J. Godoy* **and Hugo Minni** **
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral
Santa Fe, Argentina
*rjgodoy@fich.unl.edu.ar, http://purl.net/rjgodoy
**hminni@fich.unl.edu.ar

ABSTRACT

Uniform Resource Identifiers are an integral part of the current Architecture of the World Wide Web. This work analyzes the implications and possibilities of using Universal Resource Names as unique and persistent identifiers in systems for management of decentralized content and federated collections.

Particularly, discussion focuses on applying such identifiers on the context of a learning object repository that the authors are developing at Universidad Nacional del Litoral, according to the IEEE 1484.12.1 standard for Learning Object Metadata.

It is explained why Uniform Resource Locators are inadequate, and why Universal Resource Names are preferable. A standardized resolution service over Hypertext Transfer Protocol is recommended for locating resources, and usage of Uniform Resource Characteristics for accessing Learning Object Metadata is proposed. Finally, a content-negotiation mechanism for selecting the best representation among several format or language variants is outlined.

The proposed naming schema provides a double-indirection mechanism, comparable to the Human-Friendly Names approach proposed by Ballintijn, van Steen, and Tanenbaum for improving scalability and usability in naming replicated resources.

Keywords: Learning Objects, Knowledge Repositories, Identifiers, Content-Negotiation, Education Informatics.

1. INTRODUCTION

In the last years, there has been an ongoing discussion about Uniform Resource Identifiers (URIs) and their advantages in comparison with Uniform Resource Locators (URLs) [1, 2]. URIs are an integral part of the current Architecture of the World Wide Web, as well as the Semantic Web initiative [3].

"Global naming leads to global network effects (...) To benefit from and increase the value of the World Wide Web, (...) a resource should have an associated URI if another party might reasonably want to create a hypertext link to it, make or refute assertions about it, retrieve or cache a representation of it, include all or part of it by reference into another representation, annotate it, or perform other operations on it. Software developers should expect that sharing URIs across applications will be useful, even if that utility is not initially evident."[4]

The election of *unique* and *persistent* identifiers is an important matter when dealing with decentralized content management and federated collections, which are often loose constructs without significant central authority [5]. Additionally, implementing *standardized resolution methods* is indispensable for large-scale deployment and interoperability with other systems.

The authors' interest is to utilize URIs as identifiers on a Knowledge Repository they are developing, which will be used in a university educational context [6].

It must be noted that although the analysis takes place within the specific scope of Learning Object Metadata (LOM), some results may be applied to general applications that make use of URL and other identifiers.

2.BACKGROUND

In knowledge-management and storage systems intended for supporting learning the data entities are denominated Learning Objects (LOs). A LO is a resource (either digital or non-digital) which may be used for learning, education or training [7].

Metadata is required in order to describe LOs, enabling learners and instructors to search, evaluate and utilize them; and standards compliance leads to a uniform style, enhancing the possibilities of sharing, reuse, and exchange of contents. The IEEE standard for *Learning Object Metadata* (LOM) was chosen among several others because it specifies a conceptual data schema (the "*base schema*") that emphasizes on the minimal set of attributes needed to allow these LOs to be managed and located.

Naming Requirements

Each LO and each metadata instance is identified (according to the base schema) by a pair composed of a `Catalog` element, which is the name of an identification or cataloging scheme, and an `Entry` element, which is the value of the identifier itself and belongs to the given catalog. For instance, URIs may be used as identifier entries under the "`URI`" catalog; other possible catalogs include International Standard Book Number (ISBN), Library of Congress Control Number (LCCN) and ARIADNE among others.

Identifiers must be *unique* in the sense they univocally identify a resource, albeit a single resource may be identified by more than one identifier.

LOM Identifiers and URI

The URI value-space is divided in *schemes*. Each scheme defines its own mechanisms for generation and resolution of identifiers.

"A URI can be further classified as a locator, a name, or both. The term URL refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource by describing its primary access mechanism (e.g. its network "location"). The term "Universal Resource Name" has been used historically to refer to both URI under the `urn` scheme [8], which are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable, and to any other URI with the properties of a name."[9]

The `urn` scheme is further subdivided into *namespaces*, and each namespace defines additional mechanisms in order to guarantee persistence and global uniqueness. As of June 2008, 64 URI schemes and 39 formal URN namespaces have been registered [10, 11].

Some of these namespaces are only meant for identifying documents generated by a particular organization (such as "`urn:ietf:`" for the Internet Engineering Task Force [12], and "`urn:iso`" for the International Organization for Standardization [13]), while others (e.g., OID — object identifier [14]) have general purposes.

Assignment of identifiers within a URN namespace usually requires approval by a central authority, which may delegate this responsibility to others. Few namespaces do not require a registration mechanism because they make use of a unique value which have been asigned with other purpose, such as Internet Domains and IEEE 802 MAC addresses; some namespaces of this kind are: "`urn:publicid`" [15] (ISO 8879 [16] public identifiers expressed in URI syntax), "`urn:uuid`" (unique identifiers) [17] and "`urn:fdc`" (federated content identifiers) [5].

Among URI schemes and Universal Resource Name (URN) namespaces, `urn:fdc` was found to best fulfill the requirements of simple assignment and global resolution for distributed systems (though others schemes or namespaces may be used in particular cases). On the other hand, URLs are not suitable as identifiers, because they are inherently non-persistent. [18]

URN Resolution

URN resolution is the process of translating a URN into *Uniform Resource Locator* (URL) or *Uniform Resource Characteristics* (URC) [19]. Resolution services, defined in RFC 2483 [20], provide a uniform interface for performing these conversions. They are given mnemonic names, such as N2L (which stands for URN to URL), N2R (URN to resource), etc. Some services yield a single result, while others yield multiple results (e.g., *all* the locations of a resource). There also are services that carry out the inverse conversion (e.g., they gather the URNs for a given URL).

THTTP (Trivial Convention for using HTTP in URN Resolution) protocol [21] specifies how to access resolution services via traditional Hypertext Transfer Protocol (HTTP) GET requests. The services implemented by THTTP are shown in Fig. 1.

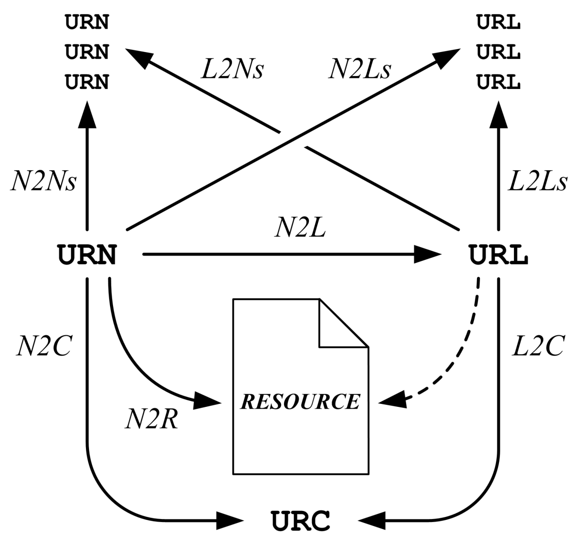

Figure 1: Identifier resolution: the solid arrows represent THTTP services, the dashed arrow represents the usual process that returns a resource given its URL.

3. PROPOSAL

Use of URN as Resource Identifiers

Using URLs as identifiers is a common practice and it has two obvious advantages: it is straightforward to get the identified resource (or a related resource thereof) given its identifier, and those resources that are accessible via HTTP or File Transfer Protocol (FTP) have already a URI of the URL kind. However, the intended semantics of URLs is to locate, not to identify, and these apparent advantages are outweighed by the advantages of using URNs.

Identifiers must be independent from the resource location and it must be possible to keep the same identifier after moving the resource. Additionally, a LO may be tagged as "unavailable", or it may be of a non-digital nature (i.e. a physical resource whose metadata is recorded in the system); in this situation it *cannot* be associated with a true URL which dereferences it.

Despite they are less common than URL, and despite of their need of namespace management, URNs are adequate for addressing these problems. Anyway, if persistence is honored and identifiers are never modified, it follows that URL-based identifiers will become outdated; and supporting deprecated or fake URLs (even though they are syntactically valid) requires as much effort as supporting identifiers that do not disclose the location.

Accessing LOM Metadata

Uniform Resource Characteristics (URC) are generic metadata about resources. They are vaguely defined in RFC 2483 as descriptions that may include "a bibliographic citation, a digital signature, or a revision history", but the content of any response to a URC request is not specified [20]. Since LO are *described* by metadata instances, it seems natural to access LOM metadata as Uniform Resource Characteristics (URC) via THTTP services N2C/L2C.

This approach provides a uniform interface for accessing LOM instances, which is similar to the resolution methods for accessing resources (N2R) or locations (N2L), thus avoiding application-specific retrieval mechanisms.

The type of URC to be returned is specified by a Multipurpose Internet Mail Extensions (MIME) [22] type, which does not only identifies the format of the result (as usual), but also its content. This requires a semantically unambiguous MIME type in order to indicate that LOM XML (Extensible Markup Language) metadata is requested, instead of other metadata (which may be optionally supported).

The MIME type `text/xml` is too general because it does not state that LOM is specifically required. A hypothetical `text/lom` type (which does not exist) would not be correct because LOM may be also encoded as Resource Description Framework (RDF) and other bindings may be defined in the future.

The `+xml` suffix [23] was defined for dealing with XML-based MIME types. For instance, some applications would be able to understand entities of `text/lom+xml` type, while others (e.g., an XML viewer) will treat them as generic XML documents. Moreover, applications without explicit support for `text/xml` will treat them as plain text.

In this case, `text/x.lom+xml` should be used because `text/lom+xml` does not exist. The `x.` prefix implies the subtype belongs to the unregistered experimental tree. (As a side note, the LOM RDF encoding cannot be expressed in the same way, because there is no `+rdf` suffix.)

LO Variants and Content Negotiation

A resource may be available in multiple representations (e.g., translations to different languages, or slides as both application/vnd.ms-powerpoint and application/pdf). Each representation is termed a *variant* of the resource. The mechanism for selecting the appropriate variant when servicing a request is known as *content negotiation* [24]. The distinction between resource and variant is a key part of the widely used HTTP protocol.

The metadata specified according to the LOM base schema include a list of languages, a list of formats and a list of locations (Fig. 2).

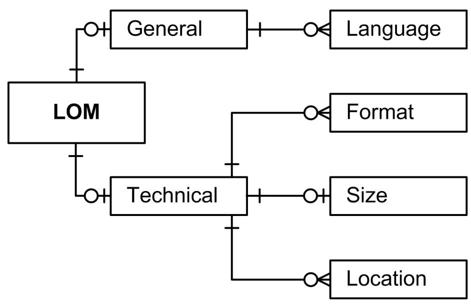

Figure 2: The elements `1.3:Language`, `4.1:Format`, `4.2:Size` and `4.3:Location` from the LOM base schema.

By analogy with HTTP, the authors had initially understood that these elements could be used for indicating different variants of a LO [25]. Under this interpretation, the information described by the LOM schema seems to be incomplete, because it would not be possible to distinguish which variant is available from each location.

While this article was being prepared, the Learning Technology Standards Committee started working on Corrigenda (not yet approved) for IEEE 1484.12.1 [26, 27]. The Committee expressly stated about this issue, noting that "*format and location are multiple for one object, which cannot have multiple sizes*". According to the correct reading of the standard, each representation constitutes a LO by itself and is related to the others by isVersionOf/hasVersion and isFormatOf/hasFormat relations. This approach is redundant, because some metadata (e.g., those about educational or pedagogic characteristics) are constant through all variants.

Since THTTP is implemented on top of the HTTP protocol, the resolution services may be accessed by general-purpose user agents (e.g., web browsers), and content negotiation may be performed via *agent-*

driven mechanisms (Fig. 3) by which the user would select a variant on the basis of several attributes from the LOM instances (e.g., language, format and technical requirements).

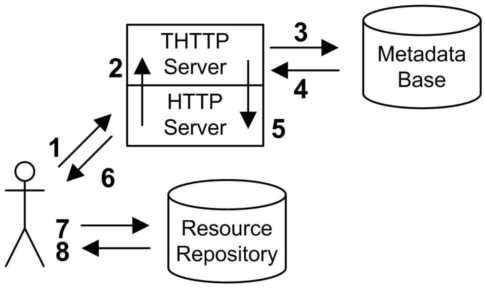

Figure 3: Agent-driven negotiation: the THTTP server receives a request (1–2), then it queries the metadata base (3–4), and returns LOM instances to the client (5–6). With this information, the client is able to identify his/her preferred variant and retrieve it from the repository (7–8).

URN for vCard externalization

Personal information about authors, editors, content providers, and other actors who contribute to the LO lifecycle, is represented in LOM as vCard 3.0 [28] entities, which are embedded into each metadata instance (Fig. 4). The authors have recommended a LOM-compliant externalization strategy (Fig.5) for storing that information in a normal form: metadata instances should contain a minimal vCard representation, and refer external vCard resources where additional (or updated) information would be located [6].

These references, indicated by means of the `source` attribute within the embedded vCard, are themselves URIs. In the original proposal `ldap:` (a URL schema) was suggested, following an example from RFC 2425 [29]. However, since the `source` attribute accepts any kind of URI, persistent identifiers (i.e. URNs) may be specified. They may be subject of the resolution mechanism explained in previous sections without introducing additional complexity to the system.

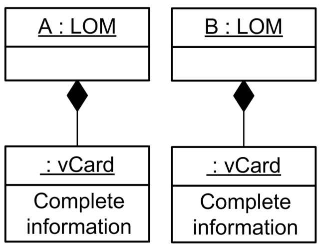

Figure 4: Two metadata instances with embedded vCard entities. The information is be duplicated if both entities refer to the same person.

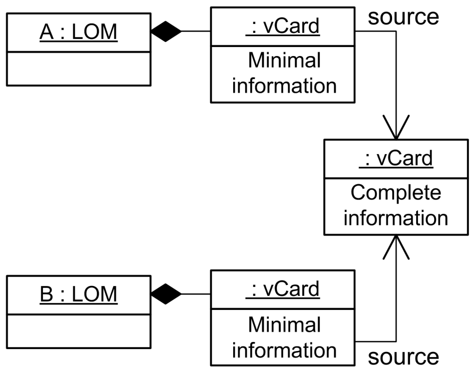

Figure 5: Two metadata instances with references to the same external vCard. The references are stored in minimal embedded vCards, in order to comply with the LOM schema.

URN as a high-level indirection layer

The LOM base schema provides a specific element (`Technical.Location`) for specifying how the contents may be accessed. This element accepts a URI as value, but this URI is intended to resolve to the content location, and not to identify the LO itself as the LOM identifiers do.

A two-step resolution process may be implemented, which is similar to the *Human-Friendly Names* (HFN) approach by Ballintijn, van Steen, and Tanenbaum [30], shown in Fig. 6. They proposed a second indirection layer, in addition to URN/URL mechanism, in order to identify resources with "names that are easy to share and remember", while URN were regarded as machine-oriented identifiers for grouping several replicas[1].

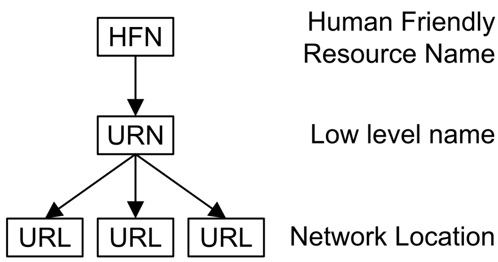

Figure 6: Naming scheme using HFN combined with a URN, as proposed in [30].

The resolution method proposed in this paper allows this kind of two-layer resolution within the scope of LOM standard: LOs are assigned with high-level human-oriented URNs, and the location of their contents is specified by other low-level URNs, as shown in Fig. 7. In turn, each low-level URN resolves to one or more URLs, which are either *mirrors* (i.e.

[1]They introduced Human-Friendly Names (HFN) as a URI scheme instead of a URN namespace. As a historical note, there was no human-oriented general-purpose URN namespaces by the time they wrote their article, but this situation has changed since then [10].

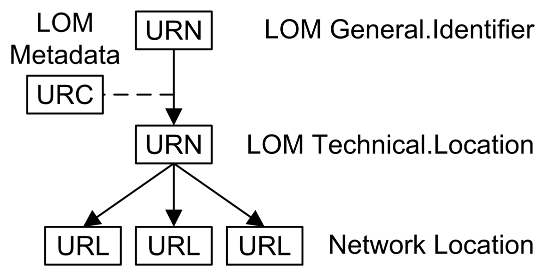

Figure 7: Two-step URN-based resolution process, accessing LOM as URC.

alternate locations) or *variants* of the resource.

4. CONCLUSION

A learning object repository is a complicated system because it must deal with granularity, versions, relations between entities, and relations between metadata and entities [31]. The complexity increases under the requirement of supporting federated collections of decentralized content.

Although there is a strong theoretical background about URN identifiers, it was found that common URL schemes are normally used, and Learning Object implementations does not take full advantage of difference between names and identifiers. (For instance, Powell et al. explicitly recommend the `http:` scheme [2, 32])

This work shows the advantages of URN in comparison with URL. URNs are preferable because they have identifier semantics and they are intrinsically persistent. In addition, several benefits from its adoption are explained.

THTTP protocol is suggested for implementing resolution services, because of three reasons:

- its implementation is very simple,
- its specification underwent enough revision as per RFC procedures [33],
- web browsers and other HTTP user agents are already enabled to access resources with no need for specialized software.

A method for encoding metadata requests by means of THTTP services is proposed, and data retrieval is enhanced with agent-driven negotiation of contents. The resolution scheme is not restricted to LO; indeed, it extends to other resources such as vCards, allowing references to personal information to be normalized according to IEEE LOM standard. This is a very important feature for the design of the repository at Universidad Nacional del Litoral, in which not only LO but also contributors are considered first class entities.

ACRONYMS

FTP	File Transfer Protocol
HFN	Human-Friendly Names [30]
HTTP	Hypertext Transfer Protocol [24]
IANA	Internet Assigned Numbers Authority
ISBN	International Standard Book Number
L2C	URL to URC a THTTP resolution service [21]
LCCN	Library of Congress Control Number
LO	Learning Object [7]
LOM	Learning Object Metadata [7]
N2L	URN to URL (a THTTP resolution service) [21]
N2R	URN to resource (a THTTP resolution service) [21]
N2C	URN to URC (a THTTP resolution service) [21]
URC	Uniform Resource Characteristics [19]
URI	Uniform Resource Identifier [9]
URN	Universal Resource Name (a URI scheme) [8]
URL	Uniform Resource Locator (a subset of URI)
MIME	Multipurpose Internet Mail Extensions [22]
THTTP	Trivial Convention for using HTTP in URN Resolution [21]
RDF	Resource Description Framework
XML	Extensible Markup Language

REFERENCES

[1] D. Booth, "URIs and the myth of resource identity," *Identity, Reference, and the Web, WWW2006 Workshop*, 2006.

[2] A. Powell, P. Johnston, L. Campbell, and P. Barker, "Guidelines for using resource identifiers in Dublin Core metadata and IEEE LOM," Dublin Core Metadata Initiative, DCMI Recommended Resource, Apr. 2005, http://www.ukoln.ac.uk/metadata/dcmi-ieee/identifiers/ (last access 2007-05).

[3] N. H. Shadbolt and T. W. Berners-Lee, "The semantic web revisited," *Intelligent Systems, IEEE*, vol. 21, no. 3, pp. 96 – 101, Feb. 2006.

[4] I. Jacobs, N. Walsh *et al.*, "Architecture of the world wide web, volume one," World Wide Web Consortium, W3C Recommendation REC-webarch-20041215, Dec. 2004.

[5] D. Tessman, "A uniform resource name (URN) namespace for federated content," Internet Engineering Task Force, RFC 4198, Nov. 2005.

[6] R. J. Godoy, H. Minni, G. Zarza, and H. Loyarte, "Design criteria for the development of an institutional learning object repository," in *Proceedings of XII Argentine Congress on Computer Science*, San Luis, 2006.

[7] Learning Technology Standards Committee, "IEEE standard for learning object metadata," Institute of Electrical and Electronics Engineers, New York, IEEE Standard 1484.12.1, 2002.

[8] R. Moats, "URN syntax," Internet Engineering Task Force, RFC 2141, May 1997.

[9] T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform resource identifier (URI): generic syntax," Internet Engineering Task Force, RFC 3986, Jan. 2005.

[10] Internet Assigned Numbers Authority, "Official IANA registry of URN namespaces," http://www.iana.org/assignments/urn-namespaces, June 2008.

[11] ——, "Uniform resource identifier (URI) schemes," http://www.iana.org/assignments/uri-schemes.html, June 2008.

[12] R. Moats, "A URN namespace for IETF documents," Internet Engineering Task Force, RFC 2648, Aug. 1999.

[13] J. Goodwin and H. Apel, "A uniform resource name (URN) namespace for the international organization for standardization (ISO)," Internet Engineering Task Force, RFC 5141, Mar. 2008.

[14] M. Mealling, "A URN namespace of object identifiers," Internet Engineering Task Force, RFC 3061, Feb. 2001.

[15] N. Walsh, J. Cowan, and P. Grosso, "A URN namespace for public identifiers," Internet Engineering Task Force, RFC 3151, Aug. 2001.

[16] International Organization for Standardization, "Information processing — text and office systems — standard generalized markup language (SGML)," International Organization for Standardization, Geneva, Switzerland, ISO Standard ISO 8879:1986(E), Oct. 1986.

[17] P. J. Leach, M. Mealling, and R. Salz, "A universally unique identifier (UUID) URN namespace," Internet Engineering Task Force, RFC 4122, Jul. 2005.

[18] R. J. Godoy and H. Minni, "Asignación y resolución de identificadores para un repositorio de objetos de aprendizaje basado en LOM," in *IX Workshop de Investigadores en Ciencias de la Computación*, Trelew, 2007, pp. pp658–662.

[19] K. Sollins, "Architectural principles of uniform resource name resolution," Internet Engineering Task Force, RFC 2276, Jan. 1998.

[20] M. Mealling and R. Daniel, "URI resolution services necessary for URN resolution," Internet Engineering Task Force, RFC 2483, Jan. 1999.

[21] R. Daniel, "A trivial convention for using HTTP in URN resolution," Internet Engineering Task Force, RFC 2169, Jun. 1997.

[22] N. Freed and N. Borenstein, "Multipurpose internet mail extensions (MIME) part two: Media types," Internet Engineering Task Force, RFC 2046, Nov. 1996.

[23] M. Murata, S. Laurent, and D. Kohn, "XML media types," Internet Engineering Task Force, RFC 3023, Jan. 2001.

[24] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext transfer protocol – HTTP/1.1," Internet Engineering Task Force, RFC 2616, Jun. 1999.

[25] R. J. Godoy and H. Minni, "Identifier management and resolution: conforming the ieee standard for learning object metadata," in *XIII Congreso Argentino de Ciencias de la Computación*, Corrientes, 2007, pp. pp967–975.

[26] E. Duval, R. J. Godoy *et al.*, "LOM v1.0 corrigenda," Learning Technology Standards Committee, Draft Minutes, 2007–2008, http://ariadne.cs.kuleuven.be/mediawiki/index.php/LOM_v1.0_Corrigenda (last access: 2008-06).

[27] Learning Technology Standards Committee, "Draft standard for learning object metadata — Corrigendum 1: Corrigenda for 1484.12.1 LOM (learning object metadata)," P1484.12.1-2002/Cor 1/D2, May 2008.

[28] F. Dawson and T. Howes, "vCard MIME directory profile," Internet Engineering Task Force, RFC 2426, Sep. 1998.

[29] T. Howes, M. Smith, and F. Dawson, "A MIME Content-Type for directory information," Internet Engineering Task Force, RFC 2425, Sep. 1998.

[30] G. Ballintijn, M. van Steen, and A. Tanenbaum, "Scalable human-friendly resource names," *Internet Computing, IEEE*, vol. 5, no. 5, pp. 20–27, Oct. 2001.

[31] C. Duncan, "Use cases for persistent identifiers," in *DCC Workshop on Persistent Identifiers*, Glasgow, Jul. 2005.

[32] A. Powell, "Identifiers - requirements and issues," in *DCC Workshop on Persistent Identifiers*, Glasgow, Jul. 2005.

[33] S. Bradner, "The internet standards process – revision 3," Internet Engineering Task Force, RFC 2026, Oct. 1996.