ARI – The Archive Resource Identifier

Chao-Chin Chou, Jan-Ming Ho
Institute of Information Science
Academia Sinica, Taiwan
{ccchou, hoho}@iis.sinica.edu.tw

ABSTRACT
In this paper, we describe the Archive Resource Identifier (ARI) System, a naming system designed to meet the requirement of the National Digital Archives Program, which aims to build digital archives of important cultural heritage and academic resources owned by major research institutes, archives, libraries and museums in Taiwan. In the ARI System, each digital resource in the archives is assigned an ARI, which is unique among all other names, and is used to manage and to fast access its instances. We also have a demonstration on the integration of the ARI System and the digital archives. The ARI System is an application of the CNRI's Handle System to the cultural heritage and academic resources in Taiwan.

Keywords
Archive Resource Identifier, Global ARI Register, Local ARI Register

1. INTRODUCTION
The National Science and Technology Program for Digital Archives, abbreviated NDAP, which is sponsored by the National Science Council (NSC), aims at building digital archives of important cultural heritage and human civilization owned by major research institutes, archives, libraries and museums in Taiwan. There are huge amounts of digital resources, such as digitized documents, photos, ancient paintings and ecological multimedia files, to be stored in separated digital archives maintained by various organizations. To efficiently manage these resources and provide an easy access mechanism for the public, a universal naming system is an urgent demand. Furthermore, the naming system should conform to our national needs, for example, it should support Chinese naming and resolution, which makes the resource names more meaningful and readable for humans. Therefore, the autonomy of the naming authorities is an essential issue that should be considered in the naming system.

There are several publicly announced naming systems. The Digital Object Identifier (DOI) System[1], developed by the International DOI Foundation, is a system for identifying and exchanging intellectual property in the digital environment. The DOI is an application of the CNRI Handle System[2] with well-defined structures on the resolution technology, metadata format and policies. However, it focuses on the intellectual-property business transactions and charges for the registration of DOIs or serves as a Registration Agency. The Content ID by the Content ID Forum (cIDf)[3], established by Professor Dr. Hiroshi Yasuda at the University of Tokyo for the purpose of providing a strong mechanism for copyright management, also gives a well-defined framework for naming digital content. However, different motives and ideas of resolution architecture between cIDf and NDAP make the Content ID inappropriate for our naming system.

The Archive Resource Identifier (ARI) System to be described is developed by the Digital Archive Task Force (DATF), which is responsible for the research and development of the technologies required by the NDAP. The ARI is intended to serve as persistent and location-independent resource identifiers for the large amount of digital resources stored in the distributed digital archives so that they can be accessed and utilized easily by the public without barriers. The ARI System defines a 2-level hierarchical framework, which allows the content holder to manage his own digital resources by his own rules, such as the naming policy, classification method and access rights, according to the characteristics of his digital resources. A general metadata format for the ARI objects is also defined in order to conveniently exchange these objects among archives and service integrators. The ARI System is also an implementation (application) of the CNRI Handle System and inherits from it the architecture, syntax of identifier and resolution technology, but is modified in accord with our national conditions and characteristics of the digital archives.

The rest of the paper is organized as follows. The syntax of ARI is described in Section 2, and its metadata format is described in Section 3. In Section 4 we give an overview of the ARI System, focusing on its hierarchical architecture of naming authorities, ARI resolution technology, and its relationships between Handle System and other naming systems. In Section 5 the integration of the ARI System and the digital archives is shown. Finally, a conclusion is drawn in Section 6.

2. ARI SYSTEM NAMESPACE

2.1 Syntax
The ARI consists of two parts: its naming authority, followed by a unique local name under the naming authority. The naming authority and the local name are separated by the ASCII character "/" (octet 0x2F). The handle syntax definition in ABNF[4] notation is shown below:

```
<ARI>::= <NamingAuthority> "/" <LocalName>

<NamingAuthority>::= *(<NamingAuthority>".") <NAsegment>

<NAsegment>::=  1*(%x00-2D / %x30-3F / %x41-FF )
                    ; any octets that map to UTF-8 encoded
                    ; Unicode 2.0 characters except octet '0x2E'
                    ; and '0x2F' which correspond to the ASCII
                    ; character '.', and '/' respectively.

<LocalName>::=  *(%x00-3F / %x41-FF )
                    ; any octets that map to UTF-8 encoded
                    ; Unicode 2.0 characters except octet '0x2F',
                    ; which corresponds to the ASCII character '/'.
```

The hierarchy of the naming authorities is implied by the combination of one or more NA segments which are separated by

the ASCII character ".". For example, "npm/ancient_painting001" is an ARI for an ancient painting in National Palace Museum, and the ARI "npm.library/ming_scroll001" stands for an ancient scroll maintained by the library of Nation Palace Museum.

The syntax of ARI inherits from the CNRI's handle syntax[5] with a slight modification in the local name. That is, we reserve the ASCII character "/" (octet 0x2F) for the reason that the local name may imply the local naming authorities of the sub namespaces when resolved by the resolution services under the Handle namespace. The details will be described in Section 4.

2.2 ARI as a Uniform Resource Name

The Uniform Resource Names (URNs)[6] are intended to serve as persistent, location-independent, resource identifiers and are designed to make it easy to map other namespaces, which share the properties of URNs, into URN-space. The ARI can also be considered as a URN, and have the syntax as follows:

URN:ari:NamingAuthority/LocalName

The URN Resolver Discovery Service (RDS)[7] framework defines a middle layer, the RDS services, which is responsible for finding out the resolution services of the sub namespaces, and different namespaces can have different resolution mechanisms, depending on their implementations and policies. Although it is still unclear on the RDS applications, the ARI Working Group of DATF is considering a formal specification as a URN. In fact, we hope that the URN RDS framework can be successful and the RDS services can be widely accepted, since it is of great importance to build localized naming systems, for example, naming systems in compliance with local cultural characteristics.

3. ARI METADATA

3.1 Overview

The ARI metadata is an essential part of the ARI System. With the metadata, the ARI is not only a location-independent identifier that maps names into URLs, but also an information provider which provides public information of the digital object it represents, including its type, qualities, attributes, owners, locations of various versions and copies, access rights, copyright statements and other alternative identifiers. The ARI metadata is recognizable by machines and provides useful information for the clawers, and hence makes the ARI more valuable in the machine-mediated network environment. Moreover, humans can format the ARI metadata according to its stylesheets to be readable via browsers. Figure 1 illustrates the ways of accessing the ARI metadata. A resolution request is handled by the ARI Resolvers, and is redirected according to the URLs specified in the ARI values to the digital archive that is responsible for the digital object. The archive then returns the ARI metadata of the requested digital object. The requestor can retrieve the instances of the digital object according to the information provided by the ARI metadata.

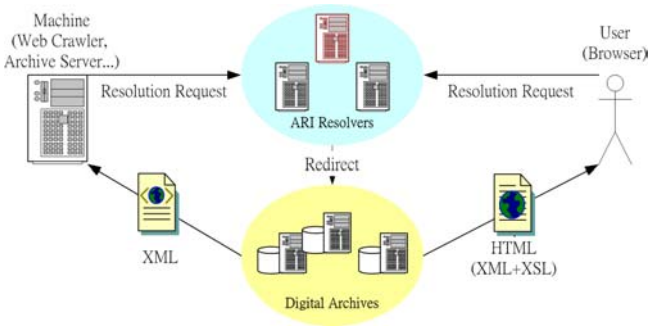

Figure 1 Accessing ARI Metadata

3.2 ARI Metadata Sets

The ARI metadata sets consist of elements of attributes, which can be classified into five categories: identifier, digital resources, ownership, access rights and locations, and are described below.

3.2.1 Attributes of Identifier

These attributes contain the information to identify the digital resource, including its ARI, its parent naming authority, the issued date, and other alternative identifiers if available such as ISBN or OID. The core elements are listed as follows:

Table 1 Core Elements of Identifier Attributes

Element	Type	Description
ARI	Required	The ARI of this digital resource.
ParentNA	Required	The parent naming authority of this digital resource.
MntInfo	Required	The management information of this ARI, including issued date, issued organization, date of last update, administrators and their contact information.
AlterID	Optional	Other available uniform identifiers of this digital resource such as ISBN or OID.

3.2.2 Attributes of Digital Resource

Elements of these attributes provide information of the digital resource, including resource type, qualities, and type specific attributes. The type specific attributes vary with the resource type. For example, for JPEG images, elements describing EXIF information are available, while for documents elements of word count and language are available. The core elements of digital resource attributes are listed as follows:

Table 2 Core Elements of Digital Resource Attributes

Element	Type	Description
Name	Required	Name or title of this digital resource.
Type	Required	Type of this resource, such as video, audio, image, document, archive, complex or naming authority.
Filetype	Optional	File type of the resource instance, such as JPEG, PDF or MP3.
Filesize	Optional	File size of the resource instance. The information is useful for the resource

		requestors to estimate the access time.
Attributes	Required	Attribute information of the digital resource. The attributes elements vary with the resource type. For example, elements for resource of type "image" consist of EXIF information, while elements for a naming authority consist of the site information.

3.2.3 Attributes of Ownership

Elements of ownership attributes provide the information of the resource owners, managers, creators and providers, about their contact information and related dates. Core elements of ownership attributes are listed as follows:

Table 3 Core Elements of Ownership Attributes

Element	Type	Description
Manager	Required	Information of the managers of the digital resource and the original content.
Owner	Required	Information of the owners of the digital resource and the original content..
Creator	Optional	Information of the creators (who created the digitized objects) of the digital resource.
Provider	Optional	Information of the providers of the digital resource and the original content.

3.2.4 Attributes of Access Rights

Elements of access rights attributes provide the information of the access rights (read, write, modify, delete) and copyrights of the digital resource. For example, some high-resolution photos of the ancient paintings have restricted copyrights and can be accessed by some researchers, and only administrators are allowed to maintain the site information of a naming authority service. Core elements of access rights are listed as follows:

Table 4 Core Elements of Access Rights Attributes

Element	Type	Description
Copyrights	Optional	The copyrights statement of the digital resource.
AccessRights	Required	Attributes of access rights, including the access lists and available access methods (access protocols).
Signature	Optional	An encrypted block to place the digital signature.

3.2.5 Attributes of Location

Elements of location attributes provide the locations of instances of the digital resource in the form of URLs. The digital resource may be presented in various forms (in video, audio, photo or plan text) and have many copies stored separately. For example, a bronze ware may be pictured from various viewpoints and with several resolutions, and the ARI that stands for the digital images of this bronze ware should provide all the locations of these photos. Core elements of location attributes are list below:

Table 5 Core Elements of Location Attributes

Element	Type	Description
URL	Optional	The URLs of the resource instances.
Filename	Optional	File name of the resource instance.

4. SYSTEM OVERVIEW

4.1 Architecture

The ARI System defines a 2-level service model. The top level is a single global resolver, the Global ARI Registry (GAR), which is unique among all other resolvers and provides the resolution and management of any ARI naming authorities. The lower level consists of all other resolvers, the Local ARI Resolvers (LARs). All the LARs should be issued and managed by the GAR as naming authority ARIs. The state information of these naming authority ARIs is the resolver information that clients can use to access and utilize associated local resolvers.

The local ARI resolver layer consists of all LARs managing all ARIs under their naming authorities, providing resolution and administration service for these local names. Local resolvers are to be hosted by the organizations that hold the digital resources, with administrative responsibility for the ARIs within the resolver or acting on behalf of the responsible organizations. Figure 2 illustrates the hierarchical architecture of the ARI System.

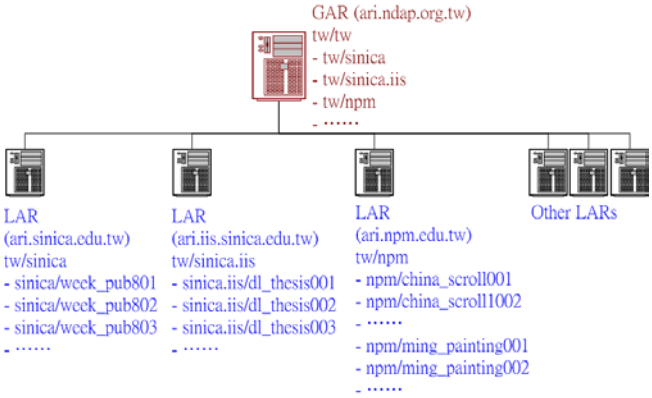

Figure 2 The 2-level hierarchies of the ARI System

4.2 ARI and Handle System

The ARI System is an implementation (application) of the Handle System to the NDAP, which is designed to conveniently exchange and access the digital resources shared by the digital archives. Although the ARI System is designed to be independent from the Handle resolution services, we ensure its compliance with the Handle System and hence the possibility to be resolved by other implementations of Handle System.

There are two ways for the identifiers in the ARI namespace to be resolved by the resolution services in the Handle namespace. One way is to make the ARI GAR and all the LARs to be registered in the CNRI GHR as local naming authorities under the Handle namespace, which is the most common way adopted by other implementations of Handle System. Figure 3 illustrates the relationship between the ARI namespace and the Handle namespace. In this case, the ARI GAR is registered in CNRI GHR as an NA handle (0.NA/1747), and has to register all the LARs to the CNRI GHR with handle names in the form of "0.NA/1747.xxx", where xxx is the local names of the LARs. All the local ARIs managed by the LARs get another identifier (the handle name) "1747.xxx/yyyy. For example, the local ARI

naming authority "tw/npm" is registered in CNRI GHR as "0.NA/1747.npm", and the ARI "npm/china-scroll1002" issued by it now has the handle name "1747.npm/china-scroll1002".

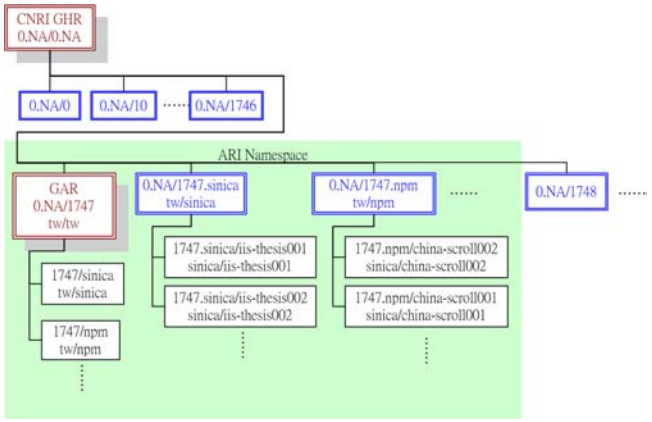

Figure 3 Resolution by the Handle System – solution 1.

The other way is to make only the ARI GAR the local naming authority under the Handle namespace. In this solution, the ARI GAR is registered in CNRI as an NA Handle, and the ARI namespace becomes a single subtree starting from this NA under the Handle namespace, as shown in Figure 4. In order to be resolved by the handle services, which do not recognize the handles issued under ARI namespace, a converted handle name is assigned to each of the ARIs, including the LARs. The converted handle names make these ARIs look like local handles issued by the API GAR (which is now a local naming authority in Handle namespace) and hence have the NA handle prefix "1747", and the hierarchies in the ARI namespace are now implied in the handle local names. For example, the ARI "npm/china-scroll1002", which is issued by the local NA "tw/npm", has the converted handle name "1747/npm/china-scroll1002". Therefore the handle server, which receives the resolution request of "1747/npm/china-scroll1002" will have it forwarded to ARI GAR according to its NA handle prefix "1747". Then the ARI GAR can parse the local name "npm/china-scroll1002", recognize that it's a local ARI issued by the registered local ARI resolver "tw/npm" (which is in fact a local handle in the GAR, i.e. "1747/npm"), and forward the resolution request of "npm/china-scroll1002" to the LAR which is responsible of the NA "tw/npm".

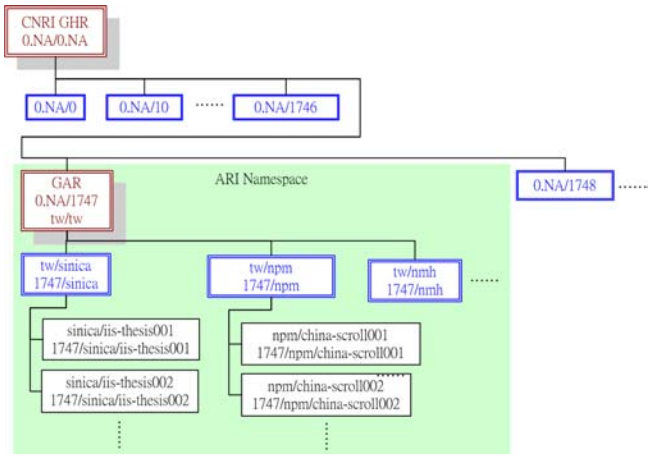

Figure 4 Resolution by the Handle System – solution 2.

5. Integration of the ARI System and Digital Archives

The integration of the ARI system and digital archives can be easily achieved by mapping each digital content to the ARI, and by

5.1 Access to the Metadata

One way to achieve the integration is to assign every digital content an unique ARI. While accessing to the metadata of a digital content from either a human user via a browser or a robot (e.g. a web crawler), the ARI resolution service can be of great help to the locating of the metadata by hiding all its physical addresses and modification history. All the metadata inquirer need to know is the ARI of the digital content he is searching for, and then through the ARI resolution service, he can get all the information needed to locate the metadata. The assignment of the ARIs to the digital contents also eases the administration of the archive for that the content managers have to keep only the data stored in the ARI system fresh. They don't have to inform all the inquirers one by one the changing or moving of their contents. Furthermore, the indexing of the digital contents by the ARIs makes the content harvesting job more efficient, because, with the resolution service, the harvester has to simply harvest all the ARIs it wants without worrying about the unnoticed location changes of the contents.

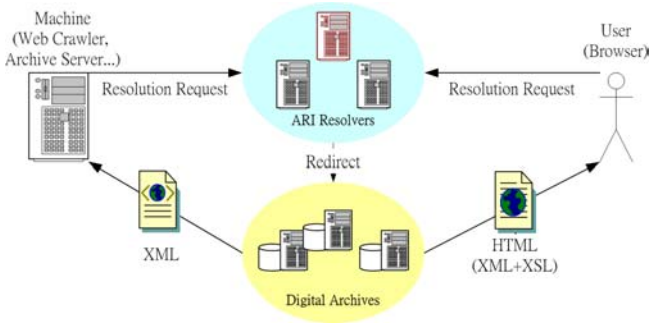

Figure 5 Access to the metadata of the digital contents via ARI resolution system.

5.2 Multimedia Files Management

There are numerous Digital Archives created in the NDAP, and all of them contains huge amount of multimedia files which includes photos, video and audio clips. The storing and processing of the multimedia files always make the content holders headache because it requires a huge storage to store these files (which usually up to hundreds of mega bytes per file), and consumes amazing CPU computing power and memories to process them. Therefore, we have proposed a centralized management service for all the multimedia files. The Multimedia Center (NDMMC), which is developed by the CSC Lab in Academia Sinica Taiwan, is designed to hold, manage, and process all the multimedia files produced by the cooperative content holders. It releases the content holders from the burden of the multimedia files management by providing them a very powerful multimedia processing system and a very huge storage with efficient management mechanism. Furthermore, the NDMMC is integrated with the ARI system by indexing all its files with ARIs. The digital archive systems can easily fetch their multimedia files by their ARIs. Therefore, all the archives systems have to keep in

their own databases are the ARIs of their multimedia files. Without the needs to store and process the multimedia files, the complexity of digital archives is greatly reduced, which also makes them more stable and allows the archive builders pay more attentions on the improving of their user interface and efficiency in content search .

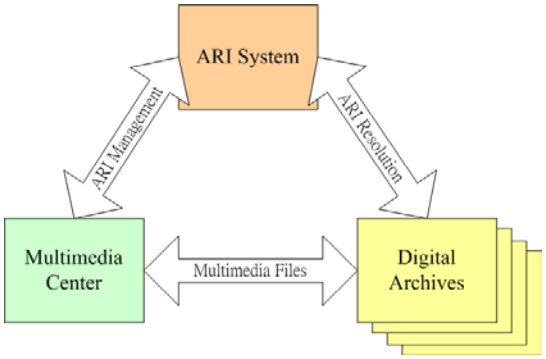

Figure 6 Integration of the ARI System and the Digital Archives

6. CONCLUSION AND FUTURE WORKS

The ARI System is a universal naming system aiming to provide both an efficient management mechanism for the content holders, and a convenient access interface for the public to the digital resources shared by the digital archives of objects of important cultural heritage owned by major research institutes, archives, libraries and museums in Taiwan. These resources are mostly provided by non-profit organizations, and are very valuable for research and education.

There are still many future works to be done for the system, such as more comprehensive naming policies and security considerations. Furthermore, more efforts should be made on removing the barrier between the general public and the resolution of ARIs. The ARI System, developed by the Digital Archive Task Force, is now a working draft toward becoming a standard for naming digital objects of the National Digital Archives Program of Taiwan and in compliance with the CNRI's Handle System.

7. ACKNOWLEDGMENTS

Thanks to Frank, Fann, Joyce, Eddie, Augcat, Jenghan Hsieh and all other team members of CSCL for giving us lots of valuable ideas and suggestions.

8. REFERENCES

[1] International DOI Foundation, Inc. "The DOI® Handbook", available at <http://www.doi.org/handbook_2000/DOIHandbookv2-5-0.pdf>

[2] Sam X. Sun, Larry Lannom, "Handle System Overview", CNRI, July 2002, available at <http://www.ietf.org/internet-drafts/draft-sun-handle-system-10.txt>

[3] The Content ID Forum, "cIDf Specification 1.1", available at <http://www.cidf.org/english/specification.html>

[4] D. Crocker, Ed., P. Overell, "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, Nov. 1997, available at <http://www.ietf.org/rfc/rfc2234.txt>

[5] Sam X. Sun, Sean Reilly, Larry Lannom, "Handle System Namespace and Service Definition", CNRI, July 2002, available at <http://www.ietf.org/internet-drafts/draft-sun-handle-system-def-07.txt>

[6] Moats, R., "URN Syntax", RFC 2141, May 1997, available at < http://www.ietf.org/rfc/rfc2141.txt >

[7] K. Sollins, MIT/LCS "Architectural Principles of Uniform Resource Name Resolution", RFC 2276, January 1998, available at < http://www.ietf.org/rfc/rfc2276.txt >

[8] Sam X. Sun, Sean Reilly, Larry Lannom and Jason Petrone, "Handle System Protocol (ver 2.1) Specification", CNRI, July 2002, available at <http://www.ietf.org/internet-drafts/draft-sun-handle-system-protocol-04.txt >