TSO
The Information Management Company

Digital Object Identifiers for Publishing and the e-Learning Community

A Report by TSO

July 2004

JISC

Digital Object Identifiers for Publishing and the e-Learning Community

A Report by TSO (The Stationery Office)

Supported by JISC funds

July 2004

JISC

Contents

1 Executive Summary and Outline Conclusions

1.1 Executive Summary

- The main purpose of this report is to guide and assist development of a JISC digital identifier policy. It is principally aimed at persistent digital identifiers for JISC services, external agencies and JISC digital content publishing programmes.
- Adoption of core digital identifier standards for data integration and interoperation would support the long-term management of JISC digital information and data. It is critical to JISC's participation in digital information communities and the growth of a distributed digital infrastructure.
- Planning a JISC digital identifier policy should be cognisant of central UK government influence on e-learning (such as DfES, Office of the e-Envoy and Becta), including government control of namespaces, government registration authorities and associated resolution services or gateways.
- The key role in JISC digital identifier planning is to define a pivotal 'Persistent Digital Identifier' framework that sets out policies for persistent data identifiers, and specifications for identifiers and associated services, to achieve interoperability and information coherence across all of the JISC's information sectors.
- Stipulating digital identifier policies, service definitions and specifications in themselves is not enough. A successful JISC implementation of persistent digital identifiers, and integration with the wider publishing community in e-learning, needs the consideration of scaled provision of: resources; customer support; best practice guidance; and a variety of centrally provided or agreed metadata schemas or specifications.
- The report recognises that it is not realistic to expect that just a single identifier system or scheme will be appropriate for, or adopted by, all of the JISC communities and activities.
- In some circumstances there are considerable advantages in being able to go beyond an opaque identifier to an 'actionable' identifier that can be resolved to single, or multiple, location(s), and, further, is able to provide authentication for access.
- It would be prudent to completely re-assess the Handle System (from CNRI), in lieu of the large provision of infrastructure for persistent identifiers, due to its widespread deployment across many publishers worldwide, and recent policy actions from governmental [53, 54], intergovernmental and worldwide e-learning information communities.

1.2 Outline Conclusions

A primary aim of JISC digital identifier planning should be to integrate and improve information services to meet the needs of the customer communities, both within the JISC environment and external to it.

Properties: Persistent digital identifiers adopted by publishers in the education sectors in the UK should be able to:

- Reference multiple object types
- Integrate and interoperate with existing standards
- Offer scope for future extension and migration
- Satisfy the needs of the broad JISC community and other sectors
- Adopt declared IETF specifications or those under known IETF approval process
- Avoid semantic or location information in the identifier

Multiple Systems: It is unlikely that the whole JISC community will be able to use a single identifier convention for the following reasons.

A single, local identifier implementation may disable medium to long-term goals of JISC information interoperability. It could create a narrow ghetto that restricts interoperability with other sectors. Thus, the broader needs of a range of other sectors should also be considered.

Many information objects or metadata records, in particular those that are provided by third-party publishers, may already have persistent digital identifiers assigned to them prior to their reference by the JISC community.

Informal sharing of information resources is likely to have different digital identifier requirements from that of the more formal traditional publishing and dissemination processes. There is a need to provide, and assure the continued availability of, more informal methods of creating persistent digital identifiers, which have low cost and minimal barriers for information providers.

Therefore, it may be essential for the JISC community to create policy and interoperation specifications to ensure that it can interact with a range of existing identifier manifestations. This would enable interoperable resource discovery, the exchange of metadata descriptions and digital objects, and aggregation relationships.

Additional Services: For some purposes, persistent digital identifiers adopted by the JISC need to do more than just provide persistent identity. They should also be able to provide extended information service functions. Dynamic actionability requires the availability of metadata and guaranteed resolution services.

The provision of JISC Identifier Services should be able to access a unique description of web-shareable resources, a principle that can be regarded as a critical key to interoperability for digital information systems across the educational sectors.

2 Introduction

2.1 Scope

This report is concerned with persistent digital identifiers and their use in the UK higher and further education sectors. The JISC Information Environment [1] is a framework of networked services and digital content that enables the discovery, access, use and publication of resources within that community. This report considers what persistent digital identifiers are, and their purpose and requirements within this environment.

The stakeholders within the JISC Information Environment who are the creators and users of such persistent digital identifiers include: publishers, authors, JISC services, researchers, tutors and learners.

Persistent digital identifiers may be associated with almost anything, but the main focus of this report is on:

- Identifiers for digital content resource for e-learning communities.

  The primary type of resource is a 'learning object', but many other diverse types of resource are also available to learners and their teachers, including: articles in journals and in institutional open archives, books, datasets, images, maps, archival collections, etc.

- Metadata records that describe these identified information resources and content objects.

2.2 Digital Identifier Definitions

Digital Identifier. Digital identifier is a generic term for a label or name composed of a sequence of characters that can be transmitted electronically. Digital identifiers can be associated with electronic, non-electronic, or abstract entities, such as books, images, reports, metadata records or events. Within this report the term 'identifier' means 'digital identifier'.

Persistence. A persistent identifier will permanently name the same resource, and will never be reused to name a different resource. This persistence refers primarily to the permanence of the identifier rather than to the resource itself. Persistence is provided through governance rather than through purely technical constraints, being the responsibility of the organisation that creates the identifier.

Namespace. A namespace is a domain in which an identifier is created and is valid.

Uniqueness. A unique identifier unambiguously names a single resource within a namespace. It will never be reused to identify a different resource.

Global and Local Identifiers. A global identifier is valid within the overall digital network in which it is used. For the JISC Information Environment, and within this report, the global digital network is the internet. A global identifier is interoperable and may be used across different systems. A local identifier is valid only within a local application or domain. A local identifier may be extended into a potentially global identifier by the inclusion of the label of its namespace. Examples of local identifiers are: a database primary key, where the namespace is the database application; a journal publisher's article acquisition number, where the namespace is the publisher's organisation.

Intelligent and Opaque Identifiers. An 'intelligent' identifier is constructed according to known rules such that: the identifier can be reconstructed by another party from knowledge of the resource; and the identified resource can be determined from studying the identifier itself. An example of an intelligent identifier is a Serial Item and Contribution Identifier (SICI). The majority of identifiers are opaque, even though some rules may be applied to their construction, and are simply strings of characters.

Collection. A collection is a set of resources, related in some way, which is treated as a single resource. Example collections are: a journal when viewed as a collection of articles; an archival collection that consists of similar items.

Functional Granularity. When assigning identifiers it is necessary to balance the level of granularity of identified items within a given content model with the functionality that will be required of the identified items. Thus, a resource is given an identifier only when there is some reason to identify it. For example:

- Should the individual diagrams or images within a learning pathway or learning object have separate identifiers? Yes, this level of use is increasingly likely and may be necessary if the copyright and the re-use of the diagrams or images are significant as virtual learning objects, with base data assets being sourced from different suppliers.
- Should individual chapters of a book have different identifiers? Yes, this would probably be useful for a book or virtual publication that is a collection of important knowledge guidance views by different authors.

Location. The location of a resource is the address at which it may be accessed. For a local resource this may be a file pathname. Within the web the location will be a URL (Universal Resource Location), probably accessible by HTTP or FTP. The location of a resource is not the same as its identifier. A resource's address may change, for example if it is moved to a different server, or if the local file structure is reorganised, whereas a persistent identifier must not change. A resource that has a single identifier may appear in more than one location, for example a journal article may be available from a publisher's website and also from an aggregator such as EBSCO.

Appropriate Copy. If an identified resource is available at several locations, a user should preferably be pointed to a version that they can access freely, possibly via a valid institution subscription. An OpenURL resolver or a handle can be used to deliver appropriate copy resolution, having knowledge of an institution's holdings and subscriptions.

Resolution and Actionable Identifiers. Resolution determines the location of a resource from its identifier. A resolver is a software application that enables a 'click' on an identifier link to yield the internet location of, and thus display, the resource, and possibly supplementary services related to the resource. Such an identifier is deemed to be 'actionable'.

Metadata. Metadata is data that describes a resource. It may be used both to discover a resource and to provide a detailed record about the resource. Metadata may include details such as the title, the author, the subject matter, the resource type and the location of a resource. It may also capture relationships between resources, such as: the identifiers of related resources; the source from which a particular resource was derived; whether a resource is part of a collection; and the version of a resource. Examples of metadata are: a library catalogue entry describing a book; the journal article header information created by a publisher that is used to provide the 'abstract' page for the article within an e-journal application. It is possible to identify some resources unambiguously by their metadata, for example a traditional journal article can be identified by the journal name, volume and issue number, and start page.

Interoperability. If metadata follows an agreed standard then it is possible for systems to exchange metadata easily, thus making the metadata interoperable.

Registry. A global registry of identifiers may provide one or both of two separate functions:

A correlation between an identifier and its location. For a persistent identifier the registry will maintain an up-to-date address for the identified resource. When providing this function the registry acts as a resolver.

A correlation between an identifier and a set of metadata that describes the identified resource. This function of a registry enables the discovery of an identifier if some details of a resource are known. Conversely, it can yield a description of an identified resource.

Within the internet there are also various registries of namespaces that assist in the definition of global identifiers. Defining identifiers within registered namespaces helps to ensure persistency and uniqueness.

Identification of Metadata Terms. For server-to-server interoperability of metadata descriptions according to different schemas, identification of metadata properties is necessary. At a meeting in late 2002, organised by the EU IST CORES project, organisations that are maintenance authorities for metadata elements reached consensus on assigning Uniform Resource Identifiers (URIs) to metadata [2]. These organisations included: GILS (Global Information Locator Service), which cross references different data sources by means of mapping; the ONIX standard book industry products, which is now standardising its metadata on the INDECS system; the Library of Congress that maintains the MARC standards for bibliographic type records; the Common European Information Research Format (CEIRF); the

IEEE Learning Object Metadata (LOM) group; and Dublin Core Metadata Initiative (DCMI). There are still issues to be resolved such as: whether metadata schemas should be built hierarchically; the problems of redundant identifiers; whether identifiers with the same meaning but different possible values should have the same name; and should different versions share the same persistent identifier. Identifiers for metadata terms are defined similarly to identifiers for resources. For example, Dublin Core metadata terms are identified within the URI HTTP namespace using PURLs (e.g. http://purl.org/dc/elements/1.1/title). Identifiers for metadata terms are outside the scope of this report and are not considered further.

2.3 Functional Requirements for Identifiers

The minimal requirements for digital identifiers within the JISC Information Environment, in particular for publishers and e-learning providers, are:

- A 'persistent' digital identifier must be capable of being registered globally, be unique and offer permanence of service over time.
- The primary purpose of a digital identifier is to name and guarantee an indirect look-up to a known object, so that it can be referred to unambiguously.
- Registered and global persistent digital identifiers must assist the accurate referencing of resources for re-use across different systems, thus enabling interoperability.
- A naming scheme within a local namespace should be independent and not imposed by any identifier registration authority. However, the JISC may wish to give some guidance on local naming schemes for JISC-sponsored resources.
- For e-learning objects it should be possible for local digital identity to be assigned by different organisations within the JISC environment independently, yet be unique.
- Locally unique resources will need to be registered for global reference if being used beyond their locale.

Further considerations that should be made when choosing and using an identification scheme are:

- The functional granularity appropriate for the identifier's intended use.
- Whether there is a need for an association between an identifier and the location of the resource. An identifier that can be resolved in some way may be actionable. Provision of a resolvable identifier probably implies the existence of a registry.
- Whether there will be multiple locations associated with a resource. This requirement places further demands on a resolution system.
- Whether there is a requirement to associate metadata describing a resource with its identifier. Provision of metadata associated with an identifier implies the existence of a registry.
- Backwards compatibility to support existing legacy naming conventions.

If metadata were to be associated with an identifier further considerations are:

- The choice of metadata scheme.
- The necessary metadata properties to capture an appropriate description of the resource. These could include:
  - the name and title of the resource
  - any alternative identifiers
  - the creator and publisher of the resource
  - the type (e.g. text, audio, video) and format of the resource
  - any related services (e.g. language translation, pay-per-view).
- Whether there is a need to record a designated authority that created and may update the metadata.
- Whether security of access to the metadata is required to respect confidentiality.
- Future extensibility of the metadata.

Processes for the management of digital identifiers need to be well defined. Some issues to consider are:

- The organisation that defines an identifier is responsible for its uniqueness and persistence and for maintaining up-to-date details of its location. Active management processes should be in place to encourage this responsibility.
- Lifecycle management of persistent digital identifiers (naming, assignment, publication, resolution and maintenance) to guarantee their global register and long-term validity.
- The investment and cost overhead in creating and maintaining central registries and systems to support publication and discovery.

2.4 Stakeholders

Within the UK further and higher education communities there are several bodies that have a stake in the application of persistent digital identifiers including:

- Traditional publishers, academic researchers, and libraries, who wish to make their content, or information about it, available electronically (e.g. e-books, journals, papers).
- Developers of content for learning materials who wish to make their e-learning materials easy to index, find, purchase and integrate within electronic learning environments.
- Educational institutions and training institutions who wish to locate and integrate educational materials into their programmes.
- Government bodies who wish to foster access to learning and are interested in opportunities offered by electronic media.
- Quality agencies who wish to maintain and enhance quality standards in new forms of learning.

The eGOV unit of the Cabinet Office, formally known as the Office of the eEnvoy, has established a set of minimum standards or specifications, the e-Government Interoperability Framework (e-GIF), to which suppliers dealing with government should conform. The part of the e-GIF that deals with learning, education and training recommends several specifications including metadata schemes, one of the mandatory elements being a unique identifier.

If the identifier for digital objects was of a type that could be resolved to multiple locations, then information could be provided in an official format where appropriate. Pilot studies are now under way by the eGOV unit of the Cabinet Office to investigate, and provide proof of concept for, the use of handle and digital object identifiers.

The E-learning Strategy Unit, in the DfES, is developing a unified e-learning strategy that includes recognition of the need for technical standards and specifications. The value of certain persistent digital identifiers, such as handles and DOIs, in supporting this strategy is being considered, particularly as ADL has recently adopted handle as the core identifier architecture for DoD in the USA [53, 54].

3 Digital Identifiers

This section reviews identifiers and metadata that are currently used to describe resources.

The requirement to be able to name or identify things is not new. Identifiers have been used to identify non-digital objects such as journals or books for some time. Knowledge and experience gained from the use of identifiers and metadata by the publishing and library communities have been taken forward into creating digital identifiers.

Digital objects include word processor documents, music and video files that may be stored on personal computers or local networks. Such a digital object has its own local filename that is assigned to it when it is created and stored. Thus it exists locally on a computer and cannot easily be found, understood or used by someone else. For individual computer users, the ability to create and name their own digital objects is essential. But if these digital objects are made publicly available (e.g. downloadable music files, training courses and product documentation) there is value in each object having a persistent, unique name.

If the digital object is available from the World Wide Web the full pathname of its file will be available globally and uniquely as a URL (uniform resource locator). But this is simply the location of the object, which may potentially change, for example if the file system of the web server is reorganised, and so it is not regarded as a valid identifier.

3.1 Uniform Resource Identifiers

A global identification for a resource in the World Wide Web is a URI (Uniform Resource Identifier) [3, 4]. There is a set of systems or namespaces for URI addresses including:

- http, e.g. http://www.bbc.co.uk/
- ftp (file transfer protocol), e.g. ftp://ftp.freeware.com
- mailto (the email addressing scheme), which could be used to identify a person or organisation, e.g. mailto: fred.bloggs@example.org

3.1.1 'http' and 'ftp' Identifiers

For a resource that is available on the web, a global identifier may be created simply within the URI http or ftp namespace, e.g. http://example.org/myresource/. However, this type of identifier provides no guarantee of persistence or resolution. Although a URI looks like a URL, so that the separation between 'name' and 'location' is not apparent, a URI is in fact an identifier that names a resource.

If the location (the URL) of the resource is the same as its URI, which is a likely scenario when the resource is first created and identified, then resolution is achieved simply via the web http. But if the resource is later moved so that its URL is changed, this straightforward resolution will fail, generally with no warning or alternative offered to a potential user. The onus is on the creator of the identifier to maintain its validity.

3.2 Universal Resource Names (URNs)

URN is a namespace of namespaces for URIs. This namespace includes:

- urn:ISBN: the namespace of book identifiers
- urn:ISSN: the namespace of serial publication identifiers
- urn:NBN: the namespace of national bibliographic numbers

These identifier schemes pre-dated the web and digital identifiers. Their namespaces were registered as URNs recognising that they are useful as digital identifiers. URN:NBN, in particular, is being used by a number of national libraries across Europe.

3.2.1 ISBN Identifiers

The International Standard Book Number System (ISBN) [5] is used to identify books or monographs, including conference proceedings. It generally identifies a particular publication so that a book may have several ISBNs. It is able to identify only the book itself so it is not suitable for identifying a part of a book such as an individual chapter. ISBNs are made up of four components, after the ISBN designator, which indicate: the group, such as area of the world or language group, of the publisher; the publisher within that group; the title of the book; and a final check digit.

An example ISBN identifier is: urn:ISBN:0262531283

3.2.2 ISSN Identifiers

The International Standard Serial Number (ISSN) is used to identify serial publications such as academic journals and periodicals. ISSN numbers are made up of eight digits, arranged in two groups of four, separated by a hyphen. An ISSN is a collection identifier because it identifies a journal and not its constituent volumes, issues and articles.

However, ISSNs are used within some journal item identifiers such as SICI. A journal may have more than one ISSN, in particular different ISSNs are assigned to the print and electronic versions. ISSNs are maintained by the ISSN agency [6]. Within the JISC Information Environment ISSN resolution, and possibly some holdings information, will be available from SUNCAT [7].

An example ISSN identifier is: urn:ISSN:0302-9743

3.2.3 NBN Identifiers

National Bibliographic Numbers (NBNs) [8] are used to identify items that have been legally deposited at a national library that fall within that library's selection policy. Each national library has a country code prefix to provide it with a namespace within which its NBNs are unique.

An example NBN identifier, from the national library of Finland, is:

urn:NBN:fi-fe19981001

3.3 Persistent URLs

A PURL (Persistent URL) [9] provides an http URI (beginning http://purl.org, and followed by the persistent identifier) that has an associated registered location with automatic resolution. When a user requests a PURL address in a web browser, the PURL registry returns the real URL to the browser. The owner of the resource must keep their entry in the PURL registry up to date so that it is always possible to locate that resource. The PURL registry is managed by OCLC.

PURLs point to only one location so they cannot permit the multiple references to different versions of a resource (e.g. Welsh, Gaelic, Gallic and English versions of government documents). They do not support metadata. PURLs are used over HTTP so may not be extensible to future web protocols that may appear. PURLs have been available for several years, but they are not widely implemented in commercial settings and they currently have a limited technical infrastructure.

An example PURL is: http://purl.org/dc/elements/1.1/title

3.3.1 PURL-based Object Identifiers

The putative PURL-based Object Identifier (POI) [10] convention is being proposed by UKOLN and OCLC to provide a localised means of assigning 'relatively' persistent global identifiers within the 'http' namespace. The convention uses the internet (DNS) domain of the organisation that owns the identifiers to provide an unambiguous, 'sub-poi' namespace. Within that namespace the organisation is then free to choose the format of its unique identifiers. The POI convention offers only 'relative' persistence as it depends on the continued use and ongoing existence of the organisation's internet domain.

Note: the POI convention is intended to provide a means of identification of a resource only. The convention does not provide resolution to location or associated metadata, the support of which would require the creation and maintenance of a registry specific to the domain.

The JISC Information Environment Service Registry (IESR) [11] is currently using POIs to good effect, identifying collections of resources (but not collection items), the services that provide access to the collections (i.e. the low-level technical access points), and the agents that own or administer the collections.

An example IESR POI is: http://purl.org/poi/iesr.ac.uk/1056366559-25788

3.4 The 'info' URI Scheme

'info' [12] is a new URI scheme that provides for registered namespaces for identifiers for bibliographic resources. It has been created because it was recognised that many of the public identifiers commonly used by the library and publishing communities are not 'official' URIs. The info URI scheme is being promoted by ANSI/NISO (the North American National Information Standards Organisation). The info URI scheme is supported by a registry of namespaces that is run by a Maintenance Agency appointed by NISO.

The info URI scheme leverages existing namespaces for bibliographic identifiers to provide global identifiers. An 'info' URI is not de-referenceable. It is simply an identifier with no associated resolution or metadata, although some of the registered schemes may provide this functionality.

Relevant namespaces currently registered within info URI are:

- oai (Open Archives Initiative identifier, described in 3.4.1) e.g. info:oai/arXiv.org:hep-th/9901001
- sici (SICI, described in 3.4.2) e.g. info:sici/07408188(200010)22:3%3C311:SEUB%3E2.0.CO;2-X
- pmid (PubMed identifier, described in 3.4.3) e.g. info:pmid/9036860
- doi (Digital Object Identifier, described in 3.7) e.g. info:doi/10.1045/july99-caplan
- hdl (Handle, described in 3.6)
- sid (Information Source identifier, based on the information provider's internet domain and an optional application identifier) e.g. info:sid/mimas.ac.uk:zetoc is a URI for the zetoc service.

Traditionally these identifiers are used without the 'info' prefix. For example, a DOI usually appears like 'doi:10.1045/july99-caplan', and this form is used in the rest of this report. Thus in situations where a global URI is required the 'info' prefix can be used.

3.4.1 Open Archives Initiative Identifier

The Open Archives Initiative [13] Protocol for Metadata Harvesting (OAI-PMH) [14] defines the required responses for Open Archives repositories to supply their metadata for harvesting by other applications. All such repositories must supply a simple Dublin Core (see 3.10.1) metadata record that describes each resource within the repository. The OAI identifier scheme [15] defines the format of identifiers for these metadata records.

An OAI identifier names the metadata record, not the actual resource and there is no required correspondence between these two identifiers. The identifier of the resource will be indicated by the value of the 'dc:identifier' property within the metadata record. The OAI identifier may be used to retrieve the associated metadata from the repository.

OAI identifiers include a namespace identifier that belongs to the organisation that provides the repository and its metadata, generally the organisation's internet domain. Within that namespace an organisation has a free choice of naming scheme to provide unique identifiers. For example, 'oai:arXiv.org:hep-th/9901001' identifies a metadata record from the arXiv open archive.

Within the JISC Information Environment there are increasing numbers of services that provide OAI-PMH metadata for harvesting, including the Resource Discovery Network (RDN) hubs, and institutional open archives of publications. Thus OAI identifiers have significant use within this environment.

3.4.2 SICI

The Serial Item and Contribution Identifier (SICI) [16] (based on an earlier BibIID system (ISO9115 1987)) is an 'intelligent' identifier for items within journals. It can identify a journal issue or an individual journal article. The SICI consists of two parts: the first part provides the unique identification of a serial issue (SII); and the optional second part a unique identifier for a serial contribution (SCI). The SII is made up of: the ISSN of the journal; the issue date; the volume and issue number; the version of the SICI standard used; and a check character. The SCI adds the contribution's location within the issue, generally the page number (and a title code in case more than one title begins on a page). The SCI, if included, appears between the issue number and the SICI version number.

For example the SICI '0363-0277(19950315)120:5<>1.0.TX;2-V' (or in its escaped form for web transmission '0363-0277(19950315)120:5%3C%3E1.0.TX;2-V') identifies the journal issue 'Library Journal, vol. 120, no. 5, March 15, 1995'.

It is possible to resolve a SICI to a location within an electronic journal application if the format of its URLs is known. SICIs effectively include their own metadata and it is possible to construct a SICI retrospectively. SICIs are suitable for the identification of serial items only.

3.4.3 PubMed Identifiers

The US National Library of Medicine provides access to abstracts of the majority of published journal articles in medical and related fields through the Entrez PubMed [17] system. Publishers provide their article header data to PubMed using a batch registration system. Information providers in the medical domain may look-up PubMed identifiers to include actionable links to abstracts. Within the JISC Information Environment PubMed identifier links may appear to end-users in applications such as references from electronic journal articles or 'abstracting and indexing' services.

3.5 XRI

OASIS [18] members have formed a new technical committee to establish a common identification scheme for distributed directory services. The Extensible Resource Identifier (XRI) [19] Technical Committee intends to create a URI scheme and a corresponding URN namespace for distributed directory services that enable the identification of resources

(including people and organisations) and the sharing of data across domains, enterprises and applications. The definition will be based on the Extensible Name Service (XNS) specifications (contributed by the XNS Public Trust Organization (XNSORG)). The committee will also define basic mechanisms for resolving XRIs, and for exchanging data and metadata associated with XRI-identified resources.

3.6 Handle Identifiers

The Handle System [20], developed by the Corporation for National Research Initiatives (CNRI) [21], is recognised as a global system for assigning, managing, and resolving persistent identifiers, known as 'handles', for digital objects and other resources on the internet.

It is a comprehensive generic internet infrastructure that offers systemisation to create persistence, location independence and global uniqueness. The Handle System is built to allow backwards compatibility with legacy systems and is capable of multiple resolutions so enabling multiple versions of the same resource to be referenced.

Identifiers are registered through a registry of naming authorities that uses a handle server. A distributed organisation that needs to use a large number of identifiers could become a naming authority by entering its location in this registry. It could then set up handle servers of its own using the freely available handle software.

Handles can be resolved in two ways. They can be resolved natively as an internet protocol through the underlying Handle System, where all handle registries offer handles look-up. Or they can be resolved by HTTP URI resolution via a web proxy. (e.g. http://www.tsoid.com or http://dx.doi.org/). A further alternative is to use a handle web client plug-in that enables the web browser to recognise the handle directly.

Handle is actively being advocated in e-learning circles [22] as a realisable persistent identifier infrastructure by a number of US federal agencies including the US Department of Defense (DoD). The recently announced ADL CORDRA development [53, 54] for SCORM 2004 is based on the use of handles as the native ID system for LOM metadata federation and cataloguing for the DoD ADL Initiative.

CORDRA will be based on existing learning content, repository and digital library standards and specifications. The focus is on selecting and adopting existing standards and specifications, and customising these if needed with community-specific profiles and extensions, not creating new standards. CORDRA will define how existing standards can be combined to meet its overarching goals.

DTIC also collaborates across federal government agencies within CENDI, a US interagency cooperative organisation, composed of the scientific and technical information managers from the Departments of Agriculture, Commerce, Energy, Education, Defence, the Environmental Protection Agency, Health and Human Services, Interior, the National Aeronautics and Space

Administration (NASA), the U.S. Government Printing Office, the National Archives and Records Administration, and the Library of Congress (the Library of Congress is an active user of handles with some 400,00 handles assigned) [23].

DTIC uses the Handle System to manage digital objects from all the US DoD agencies, and operates across the USA as a Handle Naming Authority. Handles are assigned to DTIC's full text, publicly releasable technical reports available on the internet [24]. DTIC is increasing the functionality of its handle service such that it will soon: provide secure access to unclassified but protected digital assets; support remote access and management of handles and affiliated data (as part of its partnership building effort); and extend handles to fit different digital models.

DSpace [25] is an open source software platform that enables institutions to provide open archives of their publications, including preprints and reports. It is increasingly being used by UK institutions for this purpose. DSpace uses the Handle System to assign and resolve persistent identifiers for all the digital items in a repository. The DSpace open source software package includes the handle software. Handle identifiers would be suitable for use within the JISC environment for the content of open archive repositories.

3.7 Digital Object Identifiers

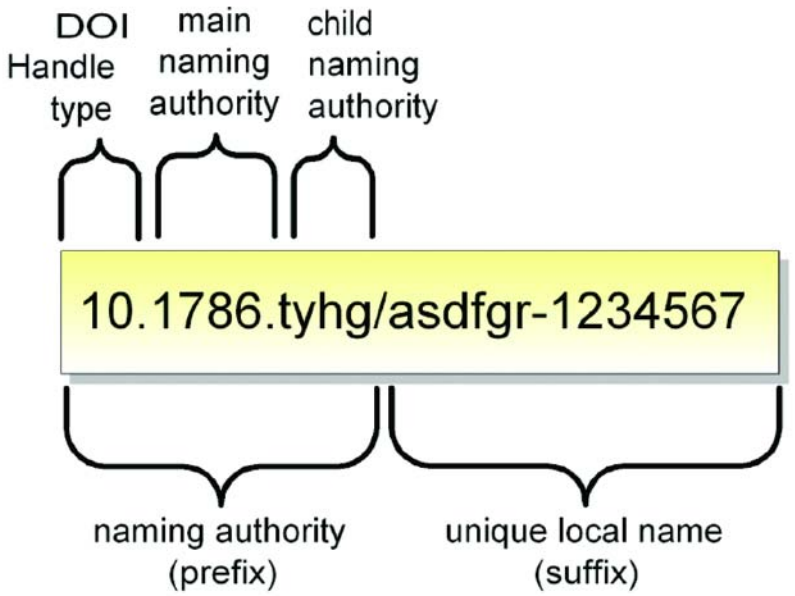
The Digital Object Identifier (DOI) System [26, 27, 28, 29] provides a framework for managing intellectual content, for linking customers with content suppliers, for facilitating electronic commerce, and enabling automated copyright management for all types of media.

The system is managed and directed by the International DOI Foundation (IDF) [30], which is a federated and not-for-profit organisation using the core internet infrastructure of the Handle System (see 3.6). Since 1997, some 13 million DOIs have now been assigned by DOI registration agencies [31] in the US, Australasia and Europe.

The syntax of a DOI is defined by a NISO standard (Z39.84-2000). It consists of a prefix and a suffix separated by a '/'. The prefix defines a namespace for the organisation that creates DOIs, known as a registration authority.

An organisation is assigned a unique prefix when it registers with a DOI registration agency. The organisation has a free choice of naming scheme for the DOI suffix.

Figure 1: Handle syntax



DOI examples:

Journal article 'Initial sequencing and analysis of the human genome', International Human Genome Sequencing Consortium, *Nature* 409, 860–921 (2001);

doi:10.1038/35057062

is based on prefix '**10.1038**' that belongs to *Nature* Magazine (the 4 digits indicate the registration agency) and a suffix '**35057062**' that was chosen by *Nature* to identify a particular article.

The office of the eEnvoy is currently using DOIs and handles to identify collections of resources related to e-GIF, the e-Government Interoperability Framework:

doi:10.1790/337224272647

Any organisation that is formally recognised as a registration authority, may register DOIs through a web form interface, supplying the identifiers and URLs that define the associated locations. There is also a batch registration process available. If the location of a resource changes, the organisation is required to update this information.

The simplest, and traditional, DOI is an identifier and its associated location.

Metadata may be associated with a DOI (see 3.10.4).

A DOI can be resolved in two ways as in the case of handle. It can be resolved natively as an internet protocol through the underlying Handle System, with all DOI registries offering handles look-up. Or they can be resolved by HTTP URI resolution via a DOI web proxy (e.g. http://www.tsoid.com or http://dx.doi.org/). A further alternative is to use a handle web client plug-in that enables the web browser to recognise the DOI directly.

The DOI System's use of the underlying handle global root infrastructure enables users who install local handle services to interoperate both with the root and with each other, depending on permissions. The DOI System adds its own dedicated expanded infrastructure, including replication servers for those IDF registration agencies that operate local handle services for their DOIs, secondary sites, mirrored servers and proxy servers housed at a secure commercial hosting service facility.

3.8 Other Identifiers

There are identification schemes available for items of specific types such as pieces of music, videos, or domestic products and comestibles. These include: the Composer, Author and Editor/Publisher of music or literary texts (CAE); the edition of printed music (International Standard Music Number – ISMN); a particular recording (International Standard Recording Code – ISRC); audiovisual works (International Standard Audio-Visual Number – ISAN); and products via the Universal Product Code/European Article Number (UPC/EAN) identifiers. These identification schemes are currently being unified in the Common Information System, into codes such as the International Standard Work Code (for musical compositions and works), to enable more efficient protection and management of copyright material.

3.9 Local Identifier Namespaces and Naming Schemes

Most of the global identifiers considered above do not require any particular naming convention for the 'local' part of an identifier. An organisation has to choose a naming scheme to ensure that identifiers are unique within their namespace.

3.9.1 Publisher Item Identifier

The Publisher Item Identifier (PII) [16] was developed by Elsevier, based on the Elsevier Standard Serial Document Identifier and the ADONIS number, when they started to publish journal articles electronically. It is used by several other publishers in the Scientific, Technical and Medical fields such as INSPEC, Springer and IEEE. The PII builds on, and incorporates, existing publisher identifiers (ISBN for books and ISSN for serial items such as journals). It is made up of three parts: a one-letter identifier (either B for books or S for serial items); then either the ISBN or the ISSN and the year of publication; followed by an identifier and a final check digit. A PII is used as a unique acquisition number and local identifier for an article throughout its life within the processes of publication. It is suitable for use as the local part of a global identifier such as the suffix of a DOI, and several publishers use a PII for this purpose. Other publishers have defined

their own naming schemes for their local part of a DOI, later in an article's lifecycle, incorporating journal volume and issue details. PIIs and related identifiers are suitable as local identifiers only for serial items and book parts, and do not have general application.

3.9.2 Organisation Namespace

Several of the identifiers listed above require an organisation's namespace as well as a local identifier unique within that namespace. The simplest namespace, without any registration requirement, but with a guarantee of uniqueness and some degree of persistence, is the organisation's internet domain (DNS). This namespace is used by OAI and POI.

Provision of such namespaces implies some form of registration, but the number of namespaces would be small. Thus a fairly informal registration process with a listing on a web page would probably be adequate.

3.9.3 A Unique Local Naming Scheme

For local identifiers defined within a small organisation, uniqueness is easy to achieve, the simplest method being sequential numbering. But within a larger organisation, or across several organisations as would be the case with creating identifiers for e-learning objects within the JISC environment, a naming scheme will be required that can be used by several people concurrently and independently. One suggested naming convention is to construct the suffix element of an identifier from the current time and the process number on the machine where the identifier is created. Both the JISC Information Environment Service Registry and the TSO DOI suffix register use this scheme to construct the local part of its identifiers, POI uses a simple Perl script. Thus in the POI identifier 'http://purl.org/poi/iesr.ac.uk/1056366559-25788':

- '1056366559' is the identifier creation time in seconds since 1970
- '25788' is the process number at creation time.

3.10 Metadata Schemes

Associating metadata with an identifier opens up wider possibilities for enhanced functionality in identifier-based applications. For practical use a registry is probably necessary to provide the correlation between the identifier and its metadata, e.g. DOI, URN:NBN.

3.10.1 Dublin Core

Dublin Core (ISO 15836) managed by the Dublin Core Metadata Initiative (DCMI) [32] provides a basic set of metadata properties to describe a resource, including its title, creators, subject coverage and description. Dublin Core's primary purpose is to enable resource discovery, hence its restriction to a simple property set. It may also be used for simple resource description but it is recognised that metadata from domain-specific schemes will be required to catalogue many resources. However, the strength of Dublin Core is its simplicity, because Dublin Core, being used and understood by many systems, provides an interoperable metadata description.

Within the JISC environment there is significant use of and requirement for Dublin Core.

Metadata descriptions provided by OAI-enabled services for metadata harvesting, and those retrieved by searching technologies used within the JISC Information Environment, are Dublin Core. The majority of more detailed metadata schemes include the simple Dublin Core properties. The UK e-Government Interoperability Framework (e-GIF) mandates a variant use of Dublin Core metadata descriptions for resources available on the web.

3.10.2 ARK

ARK is a simple set of metadata, developed by the California Digital Library registered at IETF, which focuses on the guaranteed persistence of a resource [33, 34]. This metadata format is designed both to be machine and human readable. It provides information about the resource: 'who' provided it; 'what' it is; 'where' and 'when' it was produced. ARK would provide a simple means of associating metadata with an identifier, but is not currently used within the JISC environment.

3.10.3 MARC

The MARC [35] family of metadata is a standard format used by library systems to catalogue and exchange details of their stock. It provides a standard machine-readable format in which to record metadata about bibliographic items. It is used extensively by libraries and bibliographic services such as COPAC [36]. Variations of MARC that work with current technologies, including XML, are available. MARC is one of the formats available for results retrieved by searching technologies used within the JISC Information Environment.

3.10.4 Learning Object Metadata

Learning objects should be described using specific metadata such as IMS [37] or IEEE LOM [38]. These metadata schemes are extensions of the base Dublin Core elements but include further properties to describe the pedagogical use for which the resources are intended. Within the JISC environment any metadata system associated with identifiers of e-learning objects must support learning object metadata.

3.10.5 The Digital Object Identifier Data Dictionary

All but the simplest DOIs (see 3.7) have associated metadata stored within the DOI system. There is a basic 'kernel' metadata associated with all such DOIs that indicates all the available resources associated with a particular identifier. This metadata is machine readable, thus providing the possibility of building applications that make use of it. But human-readable metadata may also be included.

In addition to the kernel metadata, DOI includes an extensive data dictionary based on the INDECS metadata framework [39, 40]. Thus, metadata appropriate to the particular type of a resource may be created. The concept of an 'application profile' is introduced, allowing customisation of the metadata set for a particular resource type, by 'mix-and-match' from the base metadata schemes.

Thus the DOI system provides three levels of DOI:

- Zero Application Profile: a simple DOI that associates a persistent identifier with a location, but has no metadata
- Base Application Profile: a DOI with associated kernel metadata
- Full Application Profile: a DOI with associated kernel metadata, plus other metadata mapped to the DOI data dictionary, plus business rules and procedures, plus any other common elements such as available services.

If a DOI has associated metadata it becomes possible to build systems that increase functionality available to end-users. An example is the CrossRef system (see 6.3).

3.11 OpenURL

A solution to the 'appropriate copy' problem is provided by the OpenURL Framework [41]. This is not actually an identifier, but is described here because it uses both identifiers and metadata for resolution. The 'OpenURL Framework for Context-Sensitive Services' is a proposed ANSI/NISO standard, Z39.88-2004 [42], which provides a way to describe a referenced resource, the 'referent', bundled together with the associated resources that comprise the context of the reference.

This package is called a 'ContextObject' and may be transported over HTTP by an OpenURL. Before the publication of the OpenURL Framework standard applications were based on the draft, 'de facto' standard, OpenURL version 0.1 [43], which transports as its 'payload' a description of the scholarly resource, such as a journal article, inline as the 'query string' of a URL.

A referent may be described by an identifier (a URI) or by metadata, as may the other entities that may be included to indicate the context. It may be encoded in either 'Key/Encoded-Value' (KEV) Format [44] or XML. KEV is a string of 'ampersand'-delimited pairs, each pair consisting of a label, or key, and an associated value, separated by an 'equals', and 'URL-encoded' for transport by HTTP.

In a digital library context, a user will click on an OpenURL link in an HTML page, for example beside a citation within a reference list of an electronic article, or alongside a record in an 'abstracting and indexing' service. The OpenURL for the reference is passed to a linking server, or 'resolver', which will return to the user a selection of resources pertinent to the cited article, preferably including a link to a copy of the full text of the article that the user is entitled to access. Typically an organisation's OpenURL resolver includes a knowledge base that records holdings, subscription and preference information specific to that organisation and decodes the metadata about the referent to determine the set of links to show to the user.

OpenURL resolvers from the major vendors are able to interwork with DOI metadata. If a referent is described by a DOI, the resolver will find the metadata about the referent from the DOI system.

The OpenURL Framework is very general and has the potential to be used in many application domains and by many communities. Its core components are defined within the OpenURL Registry [45], to provide extensibility. The Registry includes metadata formats that may be used to describe: journals and parts of journals (issues and articles); books and parts of books, including conference proceedings and papers; dissertations; and patents. There is also a simple Dublin Core metadata format available. After the standard is accepted the Registry will become available for the registration of new metadata formats.

As an example it would be possible for the e-learning communities to define and register a metadata format to support linking to e-learning objects.

OpenURL technology is providing popular appropriate copy linking to end-users within the JISC environment, mainly to journal articles. A significant number of larger institutions have purchased OpenURL resolvers. Source OpenURL links are provided by many information providers such as Web of Knowledge [46], zetoc [47], electronic journal applications and library OPACs.

Default resolvers, such as MDL LitLink [48] and ZBLSA [49] that do not have the knowledge to provide appropriate copy links, but are able to provide access to the full text of an article, are available to users at institutions that do not have an OpenURL resolver. A recently developed OpenURL router service [50] for UK academia helps information providers to link to the right OpenURL resolver for a particular user.

3.12 Summary of Global Identifiers

Standards in Table 1 are: Z39.xx – ANSI/NISO; RFCxxxx – W3C.

Table 1: Overview of Global Identifiers

	Syntax	Standard	Object type	Resolution system	Resolution type	Persistence	Metadata schema	Protocol independent	Cost	Global uniqueness	Scalability
DOI	doi or proxy URI : http://dx.doi.org	Z39.84-2000	Any	Handle & URL	multiple	y	y	y[1]	Service Costs	y	y
Handle	hdl or proxy URI : http://		Any	Handle & URL	multiple	y	x	y[2]	Service Costs	y	y
URI http	http://	RFC2616	Any	URL	single	x	x	x	x	y	y
PURL	http://purl.org/		Any	PURL Resolution service	single	y[3]	x	x	x	y	y
POI	http://purl.org/poi/		Any	x	n/a	y	x	x	Service Costs	y	y
OAI	info:oai/		Metadata	x	n/a	y	y	x	x	y	y
ISSN	Urn:ISSN:	ISO 3297	Journals	x	n/a	y	x	y	Service Costs	y	y
ISBN	Urn:ISBN:	ISO 2108	Books	x	n/a	y	x	y	Service Costs	y	y
NBN	Urn:NBN: Or proxy URI http://:	RFC3188	National Library items	Proxy URL	n/a	y	x	y	x	y	?
PubMed	info:pmid/		Abstracts (medical)	NLM Entrez	single	y	y	x	x	y	?
SICI	info:sici/	Z39.56-1996	Journal parts	x	n/a	y	y	y	x	y	y
XRI	xri:		Any	Not specified: any	multiple	y[4]	x	y	Service Costs	y[5]	y

[1] In theory it could be – but only Handle is used at the moment

[2] Handle is itself a protocol

[3] Only if the PURL–URL relationship is maintained

[4] A gradient of absolute to relative persistence can be defined

[5] Any namespace can be defined

4 Persistent Digital Object Identification Services

Sections 4 through 9 and 13 describe the digital object identifier (DOI) system as a widely used international solution to the provision of persistent global identifiers.

Providing persistent digital identifiers is more of a social and economic issue than a technical one. The implementation of a persistent identification service requires ongoing maintenance and, therefore, ongoing service resourcing. No persistent identifier service is actually free, as there is always a requirement for a reliable agency to provide services on a consistent dependable level.

At the individual agency level, any namespace and associated resolvers must be kept up to date with the current locations of the registered digital objects. The resolution provided by the system is only as up to date as the physical locations to which the persistent identifiers point. While some of this updating can be automated, responsibility for this updating, and ensuring its reliability, must be assigned within each agency, program, or office, or through a trusted third party.

It is not sufficient to create identifiers and leave them without maintenance. Active management is needed in order to gain the benefits of such a system.

4.1 Metadata for Digital Object Identifiers

For the JISC community, metadata associated with persistent identifiers could be used to: describe the resource that is associated with a given digital identifier; indicate what related services are available (e.g. language translation, pay per use etc); enable discovery of the digital identifier; and index the identifier for web search or other mechanisms.

The DOI system includes an extensive data dictionary (see 3.10.5) that allows the association of metadata, both simple and more complex, with an identifier. The resource type of an identified object is indicated by means of metadata (information about the identified resource), which is a set of optional data elements known as 'DOI kernel metadata' (see Table 2).

4.2 Building Digital Object Identifier Systems

The DOI system can offer various services, both manually operated and automated, via an identifier. From an end-user viewpoint, a link to a DOI service may be activated manually by simply clicking on it (e.g. to view a review of a particular video). The DOI system includes standardised application profiles and associated application programmer interfaces (APIs) that

enable the development of additional services. For example, such a service when linking to a particular DOI could also communicate with other DOIs. This would allow the provision of combined services with the same 'look and feel'.

For example, a student in a library may access an article in a journal to which that university library subscribes. They may then wish to access one of the articles listed in the references section of the first article. The university may not subscribe to the electronic journal containing this second journal, and it may reside on a different publisher's server. Multiple resolution services using DOIs linking with the appropriate DOI APIs, which would enable the student to access the article directly upon payment of a fee.

4.3 Minimum Requirements for Publishers and the e-Learning Community

Unique digital identifiers may be used for several information architecture purposes, e.g. to identify an object, to identify the object's metadata record, to identify a Web Resource Description Framework (RDF) metadata aggregation, etc.

In some circumstances it may be appropriate for an object to have more that one identifier associated with it, though preferably never from the same naming system. Once an identifier has been allocated it should remain unique and persistent even if the identified object ceases to exist. It is most important to observe that the management of the handles registered as DOIs in the GHS system ensures that registered identifiers are never deleted or re-used for other applications.

A key feature of DOIs, inherited from the underlying Handle System, is their ability to resolve an identifier to one or more locations. Authentication capability may also be built into this resolution. This is achieved by incorporating the digital identifier into a URI : HTTP that can be 'clicked on' in a web browser.

These locations may provide: a pointer to the object itself; some information about how it is catalogued; its version history; or some rights management functionality. The locations may change but the identifier, being a permanent reference, will remain the same.

The re-usability of digital content objects and their metadata makes it possible to distribute them widely, locate them with some form of 'search engine', and to combine them with other objects to create new content objects. For example, it would be possible to combine images from a picture library with a report, and then publish the whole item on a particular system as a new, coherent whole report with its own unique name.

There are also metadata considerations. There is widespread and increasing use of metadata records, for example in repositories and online catalogues used for describing and finding learning content. Most metadata records contain the values of a unique identifier for the described object. This could be a DOI. A single object may be useful in a range of localised contexts, and thus could have several separate metadata records.

As these records proliferate and become dispersed, the ability to determine whether they refer to manifestations of the same object becomes more important. Thus unique identification is needed, properties such as 'title' or 'location' being potentially ambiguous.

As well as the objects, the metadata records themselves may require identification, e.g. to provide tracking, or resolution to the original version or location. Metadata schemes, such as IEEE LOM, have a property that identifies the metadata record.

Digital object identifiers satisfy the requirements for persistent digital identifiers for publishers and e-learning providers, a DOI being: globally persistent and unique; independent of functional granularity; actionable, but with the possibility of multiple resolution; interoperable; potentially associated with metadata; able to record a designated authority; able to provide appropriate access to the metadata while maintaining privacy (confidentiality of metadata should be respected by those who do not require access to it); extensible; backwards compatible for legacy naming schemes; and independent, allowing a naming authority to set its own requirements for naming resources.

There should be well-defined processes for the management of digital identifiers so that they retain these properties. It would not be rational for a DOI to be made to point to two unrelated resources.

4.4 DOI System

The DOI system provides a framework for managing intellectual content, for linking customers with content suppliers, for facilitating electronic commerce, and enabling automated copyright management for all types of media. The system is managed and directed by the International DOI Foundation. Over 12 million DOIs have been assigned by DOI registration agencies in the US, Australasia and Europe.

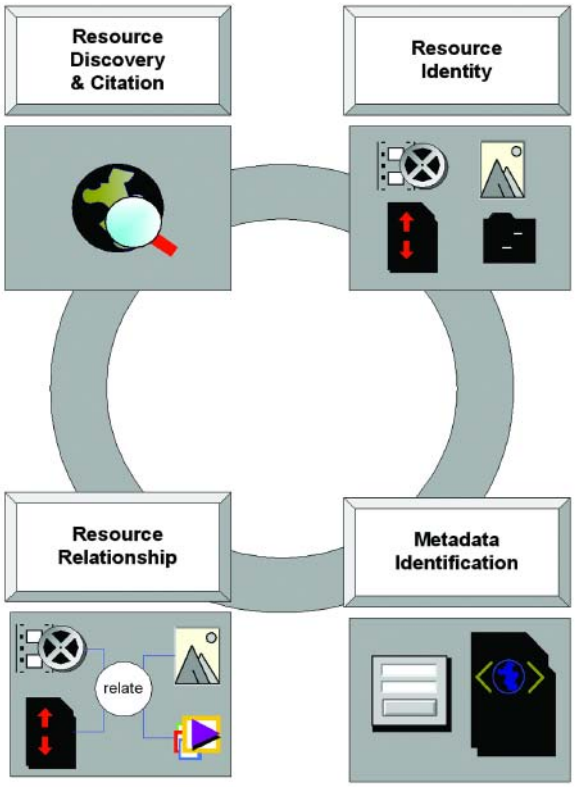
A core feature of DOIs is that they are able to offer the user a choice of meta information services of a particular resource. This facility is made possible by the inclusion of kernel metadata associated with the DOI, which indicates all the available resources associated with that particular identifier.

Once such a standard naming convention is adopted it then becomes possible not just to make resources simply available but to offer value-added services too.

4.5 Digital Object Identifiers Use

The typical use of DOIs has already pointed to some of its obvious uses, such as resource type identification, but there are other uses such as resource discovery, resource identification, metadata identification and resource relationship or syndication.

Figure 2: Use Scenarios of Digital Object Identifiers



4.6 Resource Discovery

An additional core information service, which could be connected with persistent digital identifiers, is that of resource discovery. A key reason for assigning a unique name to a resource is to enable it to be found. In the case of handle-based DOIs, this is aided by the fact that each DOI has some associated metadata making a search more precise and informative.

For this reason, it becomes possible to devise tools that look through such metadata and retrieve objects which relate to a user's aims whether that is locating a particular resource, course, learning object or other content resource.

One particularly novel discovery mode would be to perform 'reverse look-up' within cross-referenced sets of resources.

4.7 Resource Identification

Making resources available is in itself easy, but in order for these to be available publicly to a wide range of users, with different needs and goals through different systems there simply must be a way of unambiguously identifying each resource. A governed identifier, which is reliable, permanent and unique, can be used across any system now and yet is sufficiently forward facing to enable future extension, is a key element of realising this vision.

4.8 Resource Relationships: Syndication and Assembly

It would be possible to build tools to use such interlinked resources to produce a filtered list of information objects related to a key aggregation and indicate the relationships between each of the components. The development of handle/DOI-based identifiers is likely to lead to the development of new information trading exchanges. One possible area of business is in the cataloguing of available digital resources.

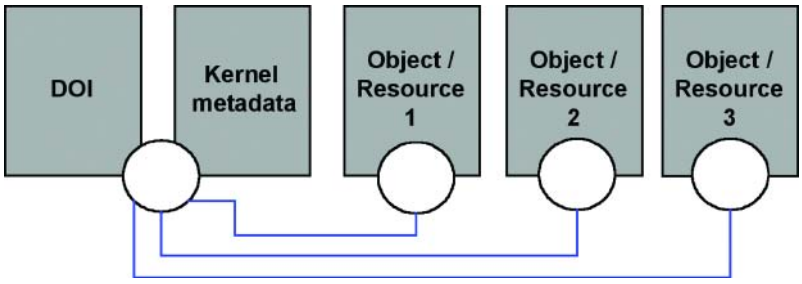
Some examples are given below:

i. Legal cases could be filtered on the basis of the legal jurisdiction in which they were based.

ii It would be useful for health inspectors to know all the available audio tracks, videos, texts and regulatory resources, which describe particular areas of food standards and quality assessment. Health inspectors could then use these resources to create virtual personalised knowledge resources in which the digital objects were the basic building blocks.

Another particular value-added service would be to create chains of linked digital objects by means of identifier kernel information that expresses the relationship between each object.

Figure 3: Service Layers for Associated Related Objects



An example of this would be to create web manifestations of each and every law and case relating to that law so that legal professionals would be able to browse either by case or by law and read up on the related material.

Such cross-referenced information would be based on metadata structures, each having their own persistent digital identifiers and could form the basis of a valuable revenue-earning service.

In order to permit such re-use of digital objects, it would be necessary to catalogue them and refer to them in that aggregation or relationship catalogue by means of their persistent DOIs. Since combining such digital resources would be time consuming, require skill and the ability to find just the right resource, it is likely that specialist aggregators would specialise in the production of such materials based on raw components.

Having such raw components available in standardised packages and being able to re-combine them would make it possible to offer a greater range of government content at a lower base cost and which would more closely meet the needs of the UK enquirer.

4.9 Digital Object Identifiers and their Kernel Metadata

A core function for DOIs is designated by means of a small base metadata set known as a 'kernel' (the critical information about each DOI). These are defined by means of a set of data elements known as 'DOI kernel metadata'.

Table 2: DOI Kernel Metadata Elements

Element (top level)	**Definition**	**Status**	**No**	**Population Value/s**
DOI	A Digital Object Id	Created automatically	1 only	DOI
DOI Genre	A class of resources with common characteristics defined by the IDF community	Created automatically	1 minimum	From Genre table
Local Identifier	A local identifier (e.g. from a legacy scheme) nominated to the resource	Optional	1 minimum	Any string but must include an identifier type, e.g. ISBN
Title	A name by which the resource is known	Optional	1 minimum	Any alphanumeric string
Type	The primary structural type of the resource	Mandatory	1 only	From: Work Abstract Physical Manifestation Digital Manifestation Performance
Origination	The process by which the resource is made	Optional	1 minimum	From: Original Derivation Excerpt Compilation Replica
Primary Agent	The name or identifier of the primary agent(s)	Created automatically	All	Identifier or Name from an agreed Genre namespace
Agent Role	The role(s) played by the primary agent(s)	Optional	1 minimum	Role from an agreed Genre namespace

The exact resource is managed at the handle record in the underlying handle technology, which enables the DOI to be resolved to a particular location at which it is stored. The handle record for the DOI is designed to accommodate very large numbers of resources and to allow full distributed administration over the internet.

4.10 Digital Object Identifier Lifecycle

This section examines the processes required for the use of DOIs in particular, though many of these lifecycle aspects are applicable to other persistent digital identifier systems.

There are five main phases: allocation, assignment, publication, resolution and maintenance.

Figure 4: Digital Object Identifier Lifecycle

4.11 Allocation

The first phase of this **simple** process concerns the reservation of a DOI/s from a registered handle naming authority or registrant (e.g. TSO) who allocates Digital Object Identifiers. What typically happens is that an individual in an organisation will first request and then create a block allocation of DOIs as un-assigned Identifiers which are not yet assigned to a resource, such as documents and PDFs. Users can create any number of DOIs at a time, up to the creation limit specified for a DOI access account. However, the permanent nature of DOIs also means that they cannot be deleted even if they are created in error. DOI blocks, which been have been requested, are then allocated to an account, which are then downloaded via an XML file. An example is shown below, normally a minimum of 500.

<DOI>10.1786/543652462725</DOI>

<DOI>10.1786/8048205431 85</DOI>

<DOI>10.1786/505650840063</DOI>

4.12 Assignment

The organisation's digital object identifier issuing agent/account holder will require the actual resolvable location of a resource (e.g. a library location identifier for a physical resource or a URL for an electronic resource) and will also need to populate metadata information about the resource, as it is likely to be made public, i.e. title, author, date of publication and its nature.

The original XML file is used to update metadata for multiple DOIs. This is achieved by uploading elements of the XML file that has been updated, back to the registrant (TSO) with the required information as metadata. This data upload is then validated and used by TSO to update the DOI handle record on the RA system. See Appendix B.

A single DOI can be updated by using the Search facility to locate the DOI in the registrant's (e.g. TSO) database. This upload information is then passed to the recognised registration authority (RA) who translates the data into their systems.

Below is an example of a DOI account record once populated and live:

- Tag > DOI: 10.1786/585587347418
- Tag > Date: 2003-11-24
- Tag > Title: Digital Identifier Pricing Information Sheet (version 1.1)
- Tag > Subject: TSO DI Pricing
- Tag > URL/Asset: http://www.tsoid.com/downloads/TSO Digital Identifier Pricing 1.1.pdf
- Tag > Creator: TSO
- Tag > Account: ********
- Tag> Public Search: Yes

4.13 Publication

Once the registration authority has entered the details of the resource into its system and the publishing organisation has placed the resource at the registered location, it then becomes possible to enter the full DOI into, for example, a web browser and gain access to the resource by a process known as resolution.

4.14 Resolution

The resolution process enables a DOI entered into a browser to be turned from an opaque 'name' into an actionable resource of use to an end-user.

Unknown to the user, the DOI is broken up into its two separate parts: the prefix which indicates the registration organisation and the organisation-determined identifier.

The system will first look at the prefix to determine the registration authority, which holds the full entry of the DOI, and it will then pass the suffix to the servers at that registration authority.

The servers at the registration authority will then return both the actual, current location of the resource and its metadata. The browser then goes to that location and retrieves the resource rather like a web page.

4.15 Maintenance

Over time it is possible that a resource pointed to by a particular reference will need to be updated. For the resource publisher this is simply a matter of refreshing the content at the location, which has been stored at the DOI registration agency that the resource publisher uses.

More subtle maintenance issues emerge in the case of one agency whose resources become the property of another agency (perhaps due to a restructuring of departments). In this instance some of the resources may be relocated to new servers and in this case the new owner simply has to inform the registration authority of the new location of the resource so that the DOI record can be updated and users of the resource will still be able to access the resource with the same DOI.

Over time some DOIs will point to obsolete resources and in this case the owner can elect to point the DOI to a helpful message rather than simply leaving the browsing tool to produce an obscure error code.

5 Costs

The actual cost provision of handle-based persistent identifiers is more of a social and economic issue than that of a technical one. The implementation of global registers and associated persistent identification resolvers requires ongoing maintenance and, therefore, ongoing permanent allocation of resources. As an example, at the individual registration authority level, the resolver must be kept up to date with the current URLs for the locations of digital objects.

The resolution services provided by the system are only as up to date as the physical locations to which the persistent identifiers point. While some of this updating can be automated, responsibility for this updating and ensuring its reliability must be assigned within each agency, program or office, or through a trusted third party.

It is not sufficient to create identifiers and leave them without maintenance; active management is needed in order to gain the benefits of such a system.

5.1 Cost of Providing a Handle System

The initial cost overhead would essentially be hardware procurement and any network overhead. There would be an ongoing maintenance operative cost, tasked to supply a reliable service for the handle data layer and associated metadata systems. (Exemplar of this architecture is depicted and described in Appendix A.)

5.1.1 Build Cost

Such initial cost is predicated on the provision of Local Handle Service's (LHS) resources (see Appendix *A.1 Handle Services and Syntax*) with development time for installation:

- Account access portal/s
- Handle LHS Server/s
- HTTP Proxy server/s
- Metadata Repository/s
- Data connections.

Given that procurement costs vary by sector and by agency type, it is improper to detail and suggest an exact costing: an average figure for the above in normal commercial terms is estimated to be £200,000 if all is procured from scratch. As a rule the JISC and services agencies will probably use existing infrastructures, which can be re-apportioned to provide the above service architecture, costs then being based on re-configuration rather than outright set-up.

5.1.2 Ongoing Annualised Cost

Cost overhead is the maintenance and provision of Local Handle Servers (LHSs), tasked to supply a reliable service for the local handle data layer and associated metadata systems.

A nominated UK government registration authority operating LHSs would be expected to provide a reliable and persistent handle management service, maintaining an orderly and consistent identifier namespace and resolver capability. This would include the:

- Account service: Staffing and resourcing
- Account management costs

Plus

- Management and maintenance of handle account access portal/s
- Management and maintenance of handle LHSs
- Management and maintenance of handle HTTP Proxy server/s
- Management and maintenance of handle Metadata Repository/s.

Again cost estimates vary, calculation being dependent on location and community of use, particularly if an organisation uses in-house resource capability and staffing. If external consultants or agencies were to be involved, estimation would be around a minimum of £190,000 per annum. This initial estimated cost would also be subject to annualised scaling, to take account of expansion of identifier volumes, the number of handles being issued, and the associated resolution service expansion.

5.2 Cost of DOI Registration Authority

This can be broken down into three cost elements:

i. Infrastructure set-up

Account web access portal/s, handle LHS Server/s, HTTP Proxy server/s, Metadata Repository/s, Data connections.

The initial registration authority build costs are to set up service for the above items, required to operate in the federated IDF Global Handle Service, but this is dependent on the specific requirement of each agency and whether they license technology from outside agencies.

ii. Cost/Fees from IDF (DOI foundation)

Annualised agreement with the IDF

This is $50,000 per annum, charged in advance on a half-yearly basis. It is for the ongoing licence cost for DOI handles, annualised maintenance of existing registered DOIs, and the IDF LHS proxy server.

DOI handle cost

The current purchase cost of an IDF licence for handles is $0.4c per unit.

The current maintenance cost per annum is $0.1c per unit.

iii. Ongoing resource costs required to manage and maintain a DOI registration authority

- Account access portal maintenance
- Handle LHS Server/s maintenance
- HTTP Proxy server/s maintenance
- Metadata Repository/s maintenance
- Content Resolution services maintenance
- Network connections and data backbone
- Account service staffing and resourcing

Estimations for the above would be around £275,000 per annum, including costs to IDF/ CNRI.

5.3 DOI Services

5.3.1 DOI End-User Costs

The end-use of digital object identifiers for users is **ZERO**. Of course the resource owner 'resolved' by the DOI may wish to charge for access to or use of the identified resource. There are no user web client requirement costs, because the DOI is expected to operate via an HTTP proxy (effectively a URL) or use a free browser handle plug-in.

5.3.2 DOI Creator/Owner Costs

The creator/owner of the DOI may need to pay registration costs and may incur additional costs from the registration agency if the resource is frequently accessed. Base DOIs (i.e. handles) are normally issued on a cost-recovery business model in which the cost is borne by the registration authority service provider of the Digital Object Identifier prefix.

The need for such charges is to recover the investment in the computer systems that the registration authorities require in order to maintain an orderly identifier namespace and resolver service. The registration authority will need to maintain sufficient bandwidth of access to its servers to ensure reliable resolution of DOIs to physical locations and have sufficient server resource to store them and maintain associated services.

The DOI owner will also incur staffing costs for the registration of new DOIs and the maintenance of the location addresses and metadata of existing DOIs.

5.4 Exemplar DOI Service Costs and Pricing: TSO

TSO Digital Object Identifier Services offers pricing policies for providing digital object identifiers and associated single resolution services that are very simple. The primary operating method for government groups is run on a cost recovery basis. That is, charges are made to meet overhead and licensing costs from the various agencies that provide TSO with core technology and services (i.e. IDF).

A PDF for the full list of TSO DOI Services can be found using the DOI 10.1786/585587347418.

TSO DOI charges are based on three distinct layers of provision.

A: TSO Digital Object Identifier Account: Annual Charges

Function: Provision of unique numbering and account manager rights

This is an annualised overhead charge for using a TSO account to create and maintain digital object identifiers. TSO provides a free unique numbering service for each account and allows up to 10 users for each digital identifier account.

B: Registration Charges

All DOI registrations for UK government and educational use are issued by the registrant TSO at **NIL** cost. On each account, DOIs are then charged a small maintenance fee (1p) to maintain the digital object identifier's persistence in the global handle register (GHS).

C: Resolution Service Charges

Function: Provision of persistent resolution services and redirection

Essentially single resolution charges are presently provided free of charge by TSO to all allocated identifiers on TSO digital object identifier accounts.

Exception: *TSO records all digital object identifier service activity and makes exception charges if a particular digital object identifier on an account becomes busy or is subject to extreme daily change i.e. if someone sets an identifier onto a popular web access point or to a very popular image/text element then TSO reserves the right to make reasonable exception charges to cover the overhead of the increased bandwidth requirement. Similarly, if large-scale changes are made to look-up data encoded on DOI numbering. In most cases TSO estimates that less than 0.01% of digital object identifiers registered will accrue exception charges.*

5.5 Resources Required for a User to Register a DOI with TSO

The minimal set of resource data required to create and register a DOI by a user are effectively what is termed 'base kernel' DOI metadata. The user needs a registered account with a DOI agency, and through the account mechanism requests a volume of unassigned DOIs. These are returned to the USER registered account email address as an XML file.

Using a manual or automated editing process offline, the account owner assigns valid information into base metadata for each DOI record in the XML file.

That file is then uploaded under a batch process back to the DOI registration agency. TSO operates an upload validation service which returns validation check data back to the account when uploads are successful.

Thus the actual resource cost for each agency using DOIs is the maintenance and provision of valid information in the submitted XML, which will be taken from the master data management on its local data store. (E.g. data tag: *the current URL assigned to a given DOI.*) This metadata cost management is affected by the total volume of assigned identifiers. In the case of TSO some 450,000 DOIs have been created, the cost to TSO being 450,000 x cost of record maintenance x annum. An example TSO DOI is:

'Health (Wales) Act 2003' : Elizabeth II. Chapter 4
Stationery Office [doi:10.1786/SIBQK6AD8RNA]

A TSO web service API is also available once the account DOI is created to allow automated update of the declared kernel metadata.

5.6 Resources Required to Maintain a DOI at a Registration Agency

Validated and assigned DOIs from user accounts are wholly maintained by the registration agency on behalf of the account owners to ensure that public look-up staging/resolution of each DOI is maintained on a 24x7 basis. In the case of TSO, this cost is provided by having a TSO DOI account. See TSO pricing information
[doi: 10.1786/585587347418]

5.7 Cost Estimation of Additional Services Around DOIs

The following are example additional services built on handle/DOI that could be offered:

- Resolution service: Reverse look-up to originating resource
- Resolution service: Multiple resolution to appropriate copy
- Resolution service: Resolution to rights management look-up

- Essentially common framework of charges is offered for

  DOI = base service + services extension.

Each extension service is broken into two cost elements

A: Cost of building extension services and ongoing maintenance

The average build cost for these could be approximated to £125,000+ (estimated entirely dependent on complexity and requirement) with ongoing maintenance estimated to 1/3 of build cost.

B: Cost charged to users to recover investment and ongoing costs

This can vary considerably by use case:

there could be small charges per DOI/handle record of £0.05 pence per unit e.g. for a permanent service of a link dependency check (using a link rot enquiry tool)

or charges could be up to £25/£50 per quarter, e.g. a quarterly charge per 'application profile DOI', for the persistent storage of multiple resolver metadata (described using RDF) which directs/maintains the aggregation manifest of a collection of distributed content references.

Any such costs are directly linked to the 'information value proposition' which the DOI content owner needs or wishes to deliver for use (see the use cases in section 8).

6 Use Case Scenarios for Digital Object Identifiers

6.1 Use Case 1

Title

Publishing With a Digital Object Identifier

Players

Digital repository

Digital object provider

Metadata author

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

Description

The simplest example of the use of a digital object identifier is the publication of a resource with a single resolution. Upon publication, the digital object identifier, is registered with a core set of metadata, and a location is provided that links via a URL to the resource itself. The publisher undertakes to the registration authority to maintain the digital object identifier metadata.

If the location remains the same there may be no need for further maintenance of the digital object identifier. If the location changes or the resource becomes unavailable the publisher submits a new URL that links to the resource, or an informative message if the resource is withdrawn. Similarly any core metadata changes are made as required.

Any references that use the digital object identifier, rather than the original location, will automatically make use of this new location so reducing the 'link rot' problem.

Exceptions

A change to the physical location of the resource requires changes to the digital object identifier record at the registration authority

Withdrawal of the resource requires a change to the digital object identifier record at the registration authority so that an error message will appear on resolution.

6.2 Use Case 2

Title

Embedding Digital Object Identifiers

Players

Digital object author

Metadata author

Digital repository

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

Description

A digital object identifier for a resource could be included with the resource itself, such as within HTML metatags, or embedded in other file formats such as Word or PDF. This digital object identifier can then provide links to one or more metadata records about the resource, possibly using an extension to the delivery software such as Acrobat.

An advantage of embedding the digital object identifier for a metadata record in the object itself is that the metadata, associated with the resource by the publisher, can be updated without the degradation. This will avoid the 'Chinese whisper' effect that can happen if various versions of metadata records associated with a resource are created and distributed. The maintenance of the metadata becomes simpler and immediately reflects any changes to the information associated with a resource, simply by using the digital object identifier embedded within it.

Exceptions

A change to the physical location of the metadata record requires changes to the digital object identifier record at the registration authority.

Withdrawal of the resource requires a change to the digital object identifier record at the registration authority so that an error message will appear on resolution.

6.3 Use Case 3

Title

Identifiers for Metadata Records

Players

Digital object author

Metadata author

Digital repository

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

The digital object identifier can resolve to the physical location of metadata

Description

A metadata record is an entity in itself. It can be useful to reference such a record and link to its location, rather than replicate and distribute the metadata record. By assigning a digital object identifier to a metadata record the relationships between the resource and the metadata can be managed more easily.

All the advantages of applying a persistent digital object identifier to a resource can also be gained by managing and uniquely identifying metadata instances.

Exceptions

When unofficial copies of resources are held then metadata will not link to these unofficial objects.

6.4 Use Case 4

Title

Digital Rights

Players

Learner

Virtual learning environment portal

Digital object author

Metadata author

Digital repository

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

The digital object identifier can resolve to the physical location of metadata

Description

A learner wishes to learn about Visual Basic and is interested in the 'Visual Basic 1' course offered through his local college. He goes to a national portal and browses to the link for the course he requires. On clicking this link he is offered several choices each with different associated costs and rights. He could download a pack of materials for learning at home, or he could gain access to the materials online via the college's virtual learning environment. He chooses the virtual learning environment option and is passed to a registration system, which takes his payment, creates a login for him and gives him access to the course.

The rights to use these course materials freely within the college environment are later purchased by the college and located on the local server, with links to associated support materials. All the course components' links are associated with the digital object identifier that resolves to digital rights information. Thus, it is only necessary to update the digital object identifier registry entry, the relocation and change of rights being transparent to end-users.

Without the use of a persistent digital object identifier all the links within all the course materials would have to have been changed by hand. The digital rights information, which either enables free use of the materials within the college, or through external payment services, needs to be changed only in one place by the publisher.

Exceptions

Use of non-persistent identifiers

Use of digital object identifiers without multiple references

6.5 Use Case 5

Title

Multiple Resolution

Players

Digital object author

Metadata author

Digital repository

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

The digital object identifier can resolve to the physical location of metadata

Description

A learner wishing to find a video clip about a particular topic goes to her preferred resource discovery service and browses the catalogue until she finds a title that appears to meet her needs. When she clicks on the link to find out more the digital object identifier link could offer her a choice of resolutions, some free reviews from other viewers of the video, some links to similar videos, and various links to merchandise applicable to that video.

After reading some of the reviews, the learner decides against that particular video but clicks on some of the links to related videos and chooses one of those. This new digital object identifier links to a range of formats, and, because she is browsing using a third-generation mobile phone, she selects the format for lower bandwidth.

This scenario is an example of the multiple resolution capability of digital object identifiers for digital objects. The first video is linked, via the metadata descriptions associated with it, to related resources. The second example video is offered in various formats applicable to different platforms. Not only does the digital object identifier automate the relationship between these related resources, it can (via linked services) control access to a resource if required. With identifiers that reference just single locations, such as PURL, this is not achievable. Moreover, as more digital objects are given identifiers it will become ever more important for them to be able to be discovered accurately (by means of associated metadata) and to be able to link to related resources. Indeed it is quite likely that much value will be gained by being able to access such cross-referenced resources. For example, it would be useful for scientific, technical and medical students to be able to list automatically and read all related papers in a particular area.

Exceptions

Use of non-persistent identifiers

Use of digital object identifiers without multiple references

6.6 Use Case 6

Title

Multiple Copies

Players

Digital object author

Metadata author

Digital repositories

Users

Secondary digital object holders

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

The digital object identifier can resolve to the physical location of metadata

Description

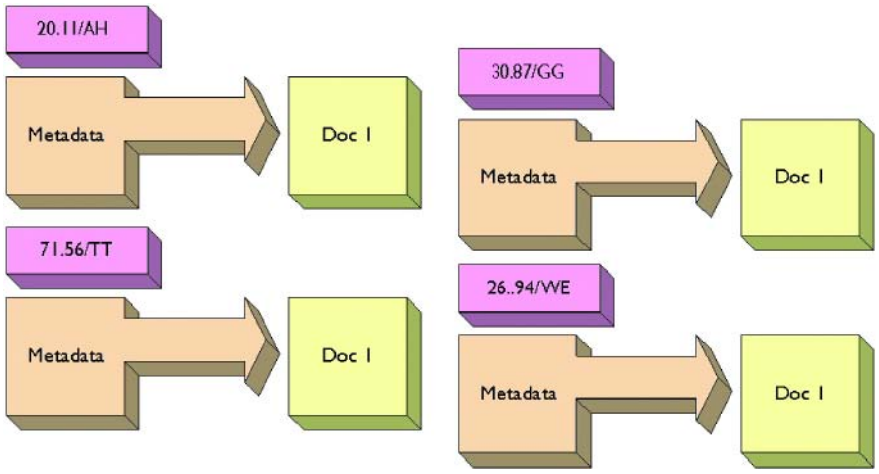
This scenario concerns a potential problem that could arise if users copy documents. Suppose an original creator of a document (for example a software manual) published it as the Digital Object Identifier 20.11/AH, along with its associated metadata and the actual document (Doc. 1). It would be possible for such a public document to be copied and assigned separate, new DOIs and metadata (30.87/GG, 71.56/TT and 26.94/WE).

Whilst each of these individual copies may well be correctly linked to the software manual they would probably each be given slightly different metadata descriptions to go with the individual DOI that had been registered against them.

Over time it will be difficult for indexing mechanisms to be able to determine which of the DOIs leads to the original document – the one which in the long term is likely to have the highest level of integrity.

What this case illustrates is that there is a need to manage both the DOI and its associated metadata. It would certainly be good practice to ensure that new versions of documents always refer to the original DOI (so that it can always be recovered) but it may be necessary to enforce integrity of data by preventing core (kernel) metadata from being copied.

Figure 5: Multiple Instances of Metadata



Exceptions

Use of non-persistent identifiers

Secure virtual learning environments in which digital objects and metadata are not copied without due access permissions

6.7 Use Case 7

Title

Cross Sector Example

Players

Lecturer

Digital object author

Metadata author

Digital repository

Curriculum Online portal

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to the URL

The digital object identifier can resolve to the physical location of metadata

Description

A publisher has produced a learning object for the schools community and submitted its metadata to Curriculum Online. This learning object is discovered by a lecturer as being useful for the further education community.

A digital object identifier was associated with this learning object by the publisher, and this can be resolved to a further digital object identifier that is associated with the metadata record that has been submitted to the Curriculum Online portal.

Examination of the digital object identifier resolutions for the learning object indicates the existence of the Curriculum Online metadata record and also resolves to digital rights metadata. It is clear that: the learning object can be purchased; who its publisher is; and how to find out more information about it.

The college content management system has a record of the resource's digital object identifier showing that the college has rights to use the learning object throughout its own intranet, but not to copy it or use it more widely.

The lecturer can access the Curriculum Online metadata record, using either the resource's or the metadata record's digital identifier in interactions with the Curriculum Online portal. From this metadata record the lecturer is able to create the basis of a new metadata record, without infringing the publisher's rights, adding information describing a use for the learning object within a further education context.

Later, the lecturer's metadata record, including the original digital object identifier for the resource, is published to a wider community and is identified using a new digital object identifier.

Links have now been established between the original object, the digital object identifier metadata from the original publisher, separate metadata records describing the learning object's use in different sector contexts, and digital rights information controlled by the publisher.

The publisher may then change the digital rights or the primary location, but the digital object identifiers provide a resolution to the original learning object and the updated rights information provided by the publisher.

Exceptions

Insecure rights management leading to a miscellany of unreliable copies

Use of digital object identifiers without multiple references

6.8 Use Case 8

Title

Researcher

Players

Library

Researcher

Digital repository

Assumptions

The repository is capable of holding content in a standard format

The objects in the repository can be assigned a URL

The digital object identifier can be resolved to multiple URLs

The digital object identifier can resolve to the physical location of metadata

Description

A researcher at a university is interested in the area of neural networks. These are artificial intelligence techniques that have drawn inspiration from biological research. This researcher might be interested in reading a paper that has been mentioned to them at a conference, with the following digital object identifier as a World Wide Web URL: http://www.publisher-doi.com/20.1986/TTY65T7E8P72H that they enter into their web browser.

The browser sees the standard hypertext reference and resolves the first part of the URL (www.publisher-doi.com), which is simply the web address of the registration authority where this Digital Object Identifier 20.1986/TTY65T7E8P72H is registered. Standard internet domain name look-up services translate this into the physical location of the registration authority's servers, to which is then passed the remainder of the digital object identifier i.e. 20.1986/TTY65T7E8P72H.

As is the convention, the first part (20.1986) of this DOI indicates its issuer, and the second part (TTY65T7E8P72H) the actual resource. The issuer is looked up at the registration authority's server, which then looks up the location of the actual resource. Finally this location is returned as a standard URL http://www.carpathia.edu/papers/mccarthy/nn65) to the web browser for display as a standard page. What is retrieved is the following page from Carpathia University:

Figure 6: Carpathia University Web Page for Professor McCarthy's Papers

Professor T. McCarthy – Papers in Computational Neuroscience	
Biological Plausibility of Recurrent Neural Networks	6.8.1.1 PDF format
Biological Plausibility of Recurrent Neural Networks	LateX format
Related Papers from Carpathia University	6.8.1.2 Related papers

This web page is actually derived from the metadata stored along with the DOI. It offers the researcher a choice of two formats in which to download the paper (PDF or LaTeX) and a link to additional related papers, which the researcher might well find relevant.

This example shows the inherent usefulness of being able to refer to more than one digital object via a single digital object identifier. There is no need to search to try to guess the correct URL at which a paper in a particular format might be stored – all available resources can be shown directly. Also there is the added value of being linked directly to further papers which are likely to be of direct relevance.

Exceptions

Use of non-persistent identifiers

Use of digital object identifiers to single references

7 Further Work

Whilst some issues are clear it became evident during writing the part of this report about digital object identifiers that further investigation is required in a number of areas. These include, but are not restricted to:

- the nature of the metadata to be used alongside adopted digital identifiers
- the processes for management of digital identifier persistence over time
- the processes for management of resource persistence over time
- details of DOIs and handle-based APIs
- enhanced services (cross referencing, rights management).

8 Recommendations

The following are recommendations for the JISC community.

- To avoid confusion, JISC should investigate, publish and disseminate a policy for persistent digital identifiers. Any such persistent digital identifiers employed should be able to identify resource objects uniquely and globally. They should also be able to incorporate metadata for some applications. Elected persistent digital identifiers should be implemented with planned flexibility and integration with other developing specifications.
- To avoid becoming a 'data ghetto' for identifiers, JISC should build services and capacity into JISC repositories to implement, and interoperate with, a range of persistent digital identifiers, such as handle, DOI or POI, even if they are not all used for JISC publications.
- It should be recognised that a multiplicity of identifiers are already in use within the JISC environment. Any future solution must encompass these existing identifiers which include: DOIs for resources from traditional publishers; handle identifiers for use by DTIC and ADL for SCORM learning object registers resources, handles within institutional Open Archives that use DSpace; POIs for collections registered with the Information Environment Service Registry (IESR); OAI identifiers for harvested metadata records; as well as traditional 'http'-based identifiers.
- JISC should now actively consider the provision of formal naming services using handle architecture. The critical issue is the creation and existence of formally known handle registers (LHS), these can be built cheaply using the UK government handle licence, as a Web Services operation (LHS implementation would require additional IT resources to develop the LHS associated registry).
- For resources or metadata records that are to be published internally on a large scale by JISC services, a recommendation is to implement a handle server with the ability to provide multiple resolution if required. Direct collaboration with ADL could be sought to integrate e-learning handle developments activity with that of the US SCORM 2004 [53] based developments (see 3.6).
- For published resources or published metadata records that are to be 'formally' published by JISC services and related to external agencies of the publishers, the report advocates wider adoption of Digital Object Identifiers as the most credible short- and medium-term solution. JISC could engage with agencies to provide short- to medium-term provision of identifier services and resolution gateways. Such cost of adoption could be minimised by the efficiency of a nominated large-scale registration authority and impacted by UK government pricing policies (UK Government Registration Agency TSO offers DOIs at zero cost as of May 2004).
- In the longer term JISC should consider becoming a DOI registration authority if the volume of registrations warrants the cost benefit of JISC core service intervention.

- The government should provide national guidance and a framework for namespace, registration authority and resolution service governance for the education sectors to support DfES goals for learning across all sectors.
- An alternative local naming scheme is given in section 3.9.3. If this were to be adopted, JISC may consider providing local namespaces within the jisc.ac.uk namespace. However, if associated location resolution and metadata were required it would be necessary to provide additional IT resources and the development of an associated registry. IESR already provides such a registry for collections within the JISC Information Environment.
- The POI convention promoted by UKOLN and OCLC should also be measured and assessed, as it represents a lower-level solution for 'relatively' persistent identifiers for e-learning, particularly for short-term use material. POIs, would need to operate in conjunction with a local naming scheme suggested in section 3.9.3. JISC may wish to consider providing local namespaces within the jisc.ac.uk namespace.
- The increasing use of OpenURL technology should be recognised, both institutional and default resolvers, to provide appropriate copy access to resources, journal articles in particular. Identifier and associated metadata solutions must be able to interoperate with OpenURL resolvers, although this requires discussion with OpenURL resolver vendors to enhance their recognition of identifier types.
- URLs, used as matching URI and location, can continue to provide an ephemeral solution as identifiers for short timescale, single location, user authored resources, with low management and minimal cost. The disadvantage is lack of persistence, but this can be countered to some extent by clear guidelines for best practice.

9 References

[1] JISC. 'Investing in the Future: Developing an Online Information Environment'. http://www.jisc.ac.uk/index.cfm?name=ie_home

[2] 'Identifying Metadata Elements with URIs', Baker, T. and Dekkers, M., *D-Lib* Magazine 9 (7–8), 2003.

[3] W3C. 'Naming and Addressing: URIs, URLs'. http://www.w3.org/Addressing/

[4] Berners-Lee, T., Fielding, R., Irvine, U.C. and Masinter, L. 'Uniform Resource Identifiers (URI): Generic Syntax'. http://www.ietf.org/rfc/rfc2396.txt

[5] The International ISBN Agency. http://www.isbn-international.org

[6] ISSN. http://www.issn.org

[7] SUNCAT: UK National Serials Union Catalogue. http://www.suncat.ac.uk

[8] Hakala, J. 'Using National Bibliography Numbers as Uniform Resource Names'. http://www.ietf.org/rfc/rfc3188.txt

[9] PURL. http://purl.oclc.org/docs/new_purl_summary.html

[10] Powell, A., Young, J., Hickey, T. 'The PURL-based Object Identifier (POI)'. http://www.ukoln.ac.uk/distributed-systems/poi/

[11] IESR: The JISC Information Environment Service Registry. http://www.mimas.ac.uk/iesr/

[12] 'info' URI Scheme. http://info-uri.info

[13] Open Archives Initiative. http://www.openarchives.org

[14] The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html

[15] Specification and XML Schema for the OAI Identifier Format. http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm

[16] Paskin, N. (1997) 'Information Identifiers'. *Learned Publishing* 10(2), pp. 135–56.

[17] National Library of Medicine. Entrez PubMed. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

[18] OASIS. http://www.oasis-open.org

[19] XRI. http://xml.coverpages.org/ni2003-01-08-a.html

[20] The Handle System. http://www.handle.net

[21] CNRI: the Corporation for National Research Initiatives. http://www.cnri.reston.va.us

[22] CETIS. ADL to make a 'repository SCORM'. http://www.cetis.ac.uk/content2/20040219153041

[23] CENDI. http://www.dtic.mil/cendi/

[24] The DTIC Handle Service. http://www.dtic.mil/dtic/handles/

[25] DSpace. http://www.dspace.org

[26] The DOI Handbook. http://www.doi.org/hb.html

[27] Paskin, N. (1999). 'DOI: Current Status and Outlook'. *D-Lib Magazine* 5(5).

[28] Paskin, N. (2002). 'Digital Object Identifiers. Information Services and Use'. 22, 97–112, IOS Press.

[29] Jacso, P. 'Digital Object Identifier'. *Information Today*.

[30] The International DOI Federation (IDF). http://www.doi.org/welcome.html

[31] DOI Registration Agencies. http://www.doi.org/registration_agencies.html

[32] The Dublin Core Metadata Initiative. http://www.dublincore.org

[33] Kunze, J.A. (2001). 'A Metadata Kernel for Electronic Permanence'. *Journal of Digital Information* 2(2).

[34] Kunze, J.A. 'The ARK Persistent Identifier Scheme'. *IETF*, 3/3/03

[35] MARC. http://www.ifla.org

[36] COPAC. http://copac.ac.uk

[37] IMS Learning Resource Meta-data Specification. http://www.imsglobal.org/metadata/index.cfm

[38] IEEE Technology Learning Standards Committee. http://ltsc.ieee.org

[39] *The Indecs Metadata framework*, Rust, G., Bide, M., 2000

[40] INDECS Summary Report, 2000

[41] Van de Sompel, H. and Beit-Arie, O. (2001). 'Open Linking in the Scholarly Environment Using the OpenURL Framework'. *D-Lib Magazine* 7(3).

[42] ANSI/NISO Z39.88-2004. 'The OpenURL Framework for Context-Sensitive Services' (Part 1: ContextObject and Transport Mechanisms, Part 2: Initial Registry Content). Available via http://library.caltech.edu/openurl/Standard.htm

[43] Van de Sompel, H., Hochstenbach, P. and Beit-Arie, O. (2000). 'OpenURL Syntax Description', Draft version 0.1. http://www.openurl.info/registry/docs/pdf/openurl-01.pdf

[44] NISO Committee AX, Apps, A. (2004). Z39.88-2004: 'The Key/Encoded-Value Format Implementation Guidelines'. http://www.openurl.info/registry/docs/implementation_guidelines

[45] Registry for the OpenURL Framework. http://www.openurl.info

[46] ISI Web of Knowledge for UK Education. http://wok.mimas.ac.uk

[47] zetoc. http://zetoc.mimas.ac.uk

[48] MDL Information Systems: LitLink. http://www.litlink.com

[49] ZBLSA. http://www.joinup.ac.uk/zblsa/

[50] The OpenURL Router. http://www.openurl.ac.uk/doc/

[51] DOI for CMS.
http://www.cmswatch.com/Features/TopicWatch/FeaturedTopic/?feature_id=66

[52] CrossRef. http://www.crossref.org
[53] ADL SCORM 2004 CORDRA Carnegie Mellon D.REHAK
http://www.lsal.cmu.edu/lsal/expertise/projects/cordra/index.html
[54] CETIS CORDRA http://www.cetis.ac.uk/content2/20040219153041

Other publications referenced during the writing of this report were:

Common Names. http://www.commonname.com

'A Uniform Resource Identifier Scheme for SNMP', Lopes, R.P. and Oliviera, J.L., IEEE Workshop on IP Operations and Management 2002

'Uniform Resource Identifiers and the Effort to Bring 'Bibliographic Control' to the Web: an Overview of Current Progress', Schwartz, R., Bulletin of the American Society for Information Science, Oct/Nov 1997, 24(1)

'Interoperability: Digital Rights Management and the Emerging Ebook Environment', Mooney, S., *D-LIB Magazine*, 2001, 7(1)

'Integration of Simultaneous Searching and Reference Linking across Bibliographic Resources on the Web', Misch, W.H., Habing, T.G. and Cole, T.W., *JCDL* 2002, ACM

'A Common Model to Support Interoperable Metadata', *D-Lib Magazine*, 5(1) 1999

10 Glossary

API	Application Programmer Interface - technical information for developing programs to work with other systems
ARK	IETF defined metadata language for describing persistent identifiers
DCMI	The Dublin Core Metadata Initiative
DOI	Digital Object Identifier
eGIF	Electronic government interoperability framework. Standards to ensure interoperable government systems
FTP	Internet File Transfer Protocol
Handle	Global look-up system infrastructure that returns the real location of a digital identifier
HTTP	Internet Hypertext Transfer Protocol
IETF	Internet Engineering Task Force - standards organisation for the internet
INDECS	EU project which defined metadata for persistent identifiers
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
JISC	The Joint Information Systems Committee of the UK further and higher education funding councils
LOM	Learning Object Metadata
MARC	Machine Readable Cataloguing - a bibliographic metadata coding scheme
NBN	National Bibliographic Number
NISO	North American National Information Standards Organisation
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OpenURL	Framework for passing information about a referenced resource and its context between servers
Persistent Identifier	A generic term for permanent identifiers such as URL, ISBN, etc.
PII	Publisher Item Identifier - persistent identifier for books
POI	PURL-based object identifier

PubMed	US National Library of Medicine identifier of medical abstracts
PURL	Persistent URL
RFC	IETF and W3C standard
SICI	Serial Item and Contribution Identifier - persistent Identifier to enable articles to be located in larger works
URI	Uniform Resource Identifier
URL	Uniform Resource Location
URN	Uniform Resource Name - URI namespace
W3C	World Wide Web Consortium
XRI	eXtensible Resource Identifier

11 Case Studies

11.1 Case Study 1: Granada Learning

Background

Granada used their allocated DOIs to reference a set of metadata records for Granada Learning products. A combination of Curriculum Online metadata and generic IMS metadata records were hosted on a test server and DOIs were created to reference each XML file. These records were accessed via a number of tools including standard web browsers, command-line tools and server-side Java web applications.

Problem Being Addressed

Persistent identification of shareable content objects (SCOs).

The Approach Taken

A selection of DOI URLs was embedded in an HTML page for testing access via a web browser. However the resources that were referenced were intended to be accessed programmatically, so this is where usage was focused. They embedded a DOI URL within a sample SCORM courseware package as an externally referenced metadata record. This was then imported into the LearnWise VLE system and the metadata verified.

The procedure for embedding the DOI URLs in HTML pages was identical to embedding any other form of URL. Granada found it extremely simple to use DOI URLs with a variety of tools.

The DOI URLs were embedded in the body of a test HTML page and as external metadata links in a SCORM courseware package manifest. The DOIs resolved successfully.

Problems Experienced

Uploading metadata was the only concern initially. This was met when the TSO XML DOI batch upload processing was explained and the template for XML exchange was set up and configured for use. In terms of content redirection, no significant performance degradation was noted when using the resolution service. Response times appeared to be as expected when accessing internet-hosted resources.

Lessons Learned

Overall Granada found working with DOIs straightforward and they will definitely consider making use of the technology in future projects. The system provides a robust framework for persistent resource location. When coupled with an API for querying and maintaining DOI embedded metadata, the technology forms a compelling solution to a range of problems.

Granada Learning sees potential use for DOIs wherever long-lived resources exist, requiring location-independent access throughout their lifespan. They believe that the technology is particularly appropriate for providing access to external resources in vendor neutral environments.

11.2 Case Study 2: Resource Discovery Network (RDN)/SOSIG

Background

The Social Science Information Gateway (SOSIG) is part of the JISC-funded UK Resource Discovery Network (RDN) and aims to provide a trusted source of selected, high-quality internet information for researchers and practitioners in the social sciences, business and law.

Problem Being Addressed

SOSIG maintains a database of more than 25,000 internet resource descriptions. Each description contains the URL of the resource being described. If an internet resource that is described by SOSIG moves to a new location, it is necessary to update the URL in the corresponding resource description. In order to build value-added services such as end-user recommended systems and ratings services around their core resource descriptions, the RDN would find it helpful to use a persistent identifier for each resource that they describe, thus shielding the underlying systems from the relatively non-persistent URLs.

As part of this case study, SOSIG have assigned DOIs for a small number of the resources that they describe, and have added the DOI to their description of each resource. Although SOSIG did not make any particular use of the assigned DOIs within their own web interface, they have exposed the DOIs in the metadata records they make available for harvesting, so that third parties can make use of them. Enough use of the DOIs was made to show that they worked and could be used in other ways.

SOSIG records are harvested into the central RDN Resource Finder database using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This central database contains copies of all RDN records and can be searched using the web interface at www.rdn.ac.uk. The DOIs were embedded in the SOSIG Dublin Core metadata records simply by adding an additional dc:identifier element.

Problems Experienced

Integration with existing systems – modifications to software were required.

Choice of DOI format.

Although the embedding was very simple, SOSIG staff had to alter a script that exports the records to OAI format as previously they had exported only the first URL in each record. Similarly, the software used to generate the 'More information' page that is part of the central RDN web interface (see Figure 7) had to be modified to look at all the available dc:identifier elements and to treat any that contained a DOI as a special case (i.e. by generating the DOI button).

There are several ways in which DOIs can be encoded as URIs. SOSIG chose to encode each DOI in their metadata records using the 'http://dx.doi.org/' form of URL. An alternative form of encoding would have used the 'doi:' URI scheme – this form is recommended in the DOI Handbook. However, SOSIG made the pragmatic decision to use the URL form so that the DOI in the metadata record would work more readily in end-user browsers.

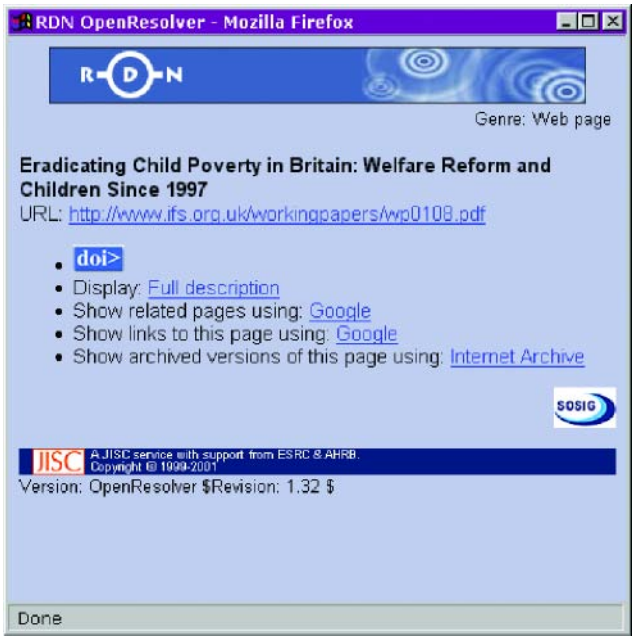
Lessons Learned

This case study has shown that DOIs can be embedded into RDN Dublin Core metadata records and shared between RDN partners (and others) using OAI-PMH. No major technical difficulties were encountered in undertaking this activity. The performance of the DOI resolver service was perceived to be acceptable.

Only a small number of DOIs were assigned. Therefore, it is not clear whether the processes (i.e. the workflow) by which DOIs are assigned and subsequently updated are sufficiently streamlined for them to be embedded into RDN workflows in an efficient and sustainable way. Nor has it given the RDN a good feel for the costs (in terms of staff effort) that might be involved in assigning and maintaining DOIs for all the resources described by the RDN.

Although this case study has not built any real end-user services on top of the assigned DOIs, it has shown that DOIs could form the basis for such services.

Figure 7: RDN "More information" window, showing DOI button



11.3 Case Study 3: TSO (The Stationery Office) Official Publications Bibliography

Background

UK Official Publications (UKOP) is the comprehensive catalogue of UK official publications and is recognised as 'the official catalogue' by Government. UKOP includes parliamentary publications, legislation – including Acts and Statutory Instruments, the publications of central government departments and devolved bodies and the output of quangos, agencies and selected international bodies, such as the UN, the WHO and the European Commission.

The UKOP catalogue contains over 450,000 records. It combines the entire TSO publications catalogue together with COBOP (The Catalogue of British Official Publications that are not published by TSO).

Problem Being Addressed

As all government departments are mandated to provide copies of their non-TSO publications for cataloguing within COBOP, the comprehensive extent of the coverage is guaranteed. TSO wanted to improve the ease of identifying and locating information within UKOP.

The Approach Taken

TSO has applied persistent digital object identifiers (DOIs) to the entire UKOP database of 450,000 metadata records. The DOIs and associated metadata records were created through an automated process using the TSO DOI API which is directly connected into the DOI/Handle System global infrastructure.

Existing metadata from the UKOP database was mapped to the DOI metadata and the DOIs were inserted by appending in to the UKOP identifier element of the record. The process took, on average, two seconds per DOI. Thus the existing UKOP metadata remained largely undisturbed, but yet crucially is linked to a DOI record with all the attendant benefits of a persistent DOI.

Examples of assigned DOIs in UKOP

'Entrepreneurship and local economic development': programme and policy recommendations /OECD [doi:10.1786/VGASVCDQR3DJ]

'Lords amendments to the Electricity (Miscellaneous Provisions) Bill' / Stationery Office [doi:10.1786/H19J5SPFTNI4]

'Statistics on smoking cessation services in England: April 2001 to March 2002' /

Department of Health [doi:10.1786/X02M6202XARG]

'House of Commons official report' / Stationery Office [doi:10.1786/G20GJVFTH65U]

Problems Experienced

As the DOI/Handle system is optimised for resolution, accessing any of the 450,000 UKOP digital object identifiers is efficient and fast. It has not resulted in any degradation to the existing service to UKOP users.

As UKOP is a subscription-based service, TSO is in the process of developing a permissions layer. Although UKOP is a subscription service, using the DOI system and architecture allows a permission layer to be used to control access to any of the exposed DOI/UKOP metadata records.

Lesson Learned

Assigning digital object identifiers to the UKOP metadata has not only improved the discoverability of the information within the database, but also created citation opportunities for much wider resource discovery, resolving and linking to other associated information.

TSO is now integrating to create a GOOGLE for government online activity, and so are continuing to make UKOP the essential online resource for public, academic and corporate organisations.

11.4 Case Study 4: University of Sunderland

Background

The University of Sunderland used DOIs to reference a range of typical objects with which higher education (HE) institutions deal. These includ ed:

- Degree course entries in online prospectuses e.g. New Route PhD, BSc and MSc courses
- Electronic course materials, course components, handouts
- Publicity material: prospectus
- Research outputs: theses, papers
- Student information: careers, library, staff, alumni
- Staff information: phone directory, funding agencies
- Physical objects: books.

Problem Being Addressed

Determining the level of awareness of persistent identifiers in HE

Determining the need for persistence of published outputs from various business areas in universities

The Approach Taken

The items selected covered a variety of different user groups from staff publishing course materials and searching for documents on the intranet, through prospective and existing student information, to administration and corporate management. Types of users were identified (lecturers, marketing, graduate school, finance, recruitment, information services, website developers).

The DOIs for the items for these user groups were embedded in a web page to test access and user reactions.

A particular person was identified and a partial re-implementation of their website was undertaken to better gauge the role of persistent identifiers.

Problems Experienced

Not only was it technically easy to embed the DOIs in the existing web-based systems and environment but it was also easy for end-users to access the objects pointed to by means of their standard web tools. The technology is therefore easy to employ and to use yet offers a number of advantages over existing technology for the HE sector:

- Stable identification across the intranet and internet
- Increased likelihood of published resources being easily discoverable
- Built-in control over access to given resources
- Enhanced services built on DOI-referenced resources.
- Some general comments from the small number of persons in typical roles in an HE institution were, however, revealing:
- Opaque DOIs are less memorable than URLs

Very few HE stakeholders have heard of DOIs (2–8% or less of those surveyed in a small study) and few that have think that such persistent identifiers will have a major impact in the short term. Students as well as staff are still wedded to paper.

Accounting and financial information needs to be maintained in a reliable archive and DOIs could contribute to this in terms of providing persistent identification.

Lesson Learned

Much groundwork still needs to be done in terms of raising awareness of the needs for persistent identification in the HE sector, what identifiers are available and how best to use them.

Cost is a concern to the higher education sector.

11.5 Case Study 5: Healthier Nation Project for FE and HE Learning

Background

The aim of the Healthier Nation project is to identify suitable learning materials from JISC-funded resource databases and/or content repositories and to re-purpose a sample of the material as learning chunks to support Health and Social Care curriculum delivery (at FE and HE levels). The project has specifically focused on the 'Big Four' diseases affecting Scotland: cancer, coronary heart disease, stroke and mental illness.

Problem Being Addressed

The objectives of the project so far have been to identify and evaluate relevant materials on the 'Big Four' diseases by:

- Carrying out an electronic search of defined and agreed databases and content repositories
- Evaluating the resources for appropriateness to health care related programmes at Scottish Credit and Qualifications Framework (SCQF) levels 4 to 7
- Identifying keywords and signposting information (levels, courses, delivery options) to facilitate the development of metadata
- Highlighting areas where little or no learning material exists to support learning in the identified subject areas at the appropriate levels.

The Approach Taken

Academic experts from each partner institution mapped resources by means of a characterisation based on key subject areas, specific keywords and exclusions, courses which could use the material (including the level) and any relevant learning outcomes. The resources were primarily taken from the relevant subject gateways, rather than bibliographic sources and indexes to increase immediate usability and potential for re-purposing.

The researcher team evaluated all the learning material on their content (clarity, authority, bias, level), style and functionality. Copyright details were also recorded for future re-purposing. Restricted vocabularies were used whenever possible, to assist metadata tagging of learning objects and academic staff were involved to evaluate appropriate use for teaching.

Problems Experienced

- The lack of agreed criteria for accessibility evaluations
- Difficulty in locating materials in RDN gateways
- Lack of suitable FE materials.

Lessons Learned

- The need to develop a model for evaluating resources and the accessibility evaluation of the resources
- Metadata development for discovering resources.

12 Acknowledgements

Editors

Robin Wilson
Director
Digital Identifier Services, TSO (The Stationery Office) Ltd

Helen Hockx
Programme Manager
JISC

Authors

Ann Apps
MIMAS, University of Manchester

Chris Bowerman
University of Sunderland

Mike Collett
Schemeta Ltd

Contributors

Prof. Benedict du Boulay
University of Sussex

Dr Steve Jeyes
Edexcel International

Charlie Jones
Granada Learning Ltd

Dr Rose Luckin
University of Sussex

Andy Powell
UKOLN, University of Bath

Alan Teece
Granada Learning Ltd

Ann Apps
MIMAS, University of Manchester

Larry Lannom
CNRI
Washington, USA

Marc Stephenson
TSO Ltd

13 Appendix A: The Handle Architecture

A.1 Handle Services and Syntax

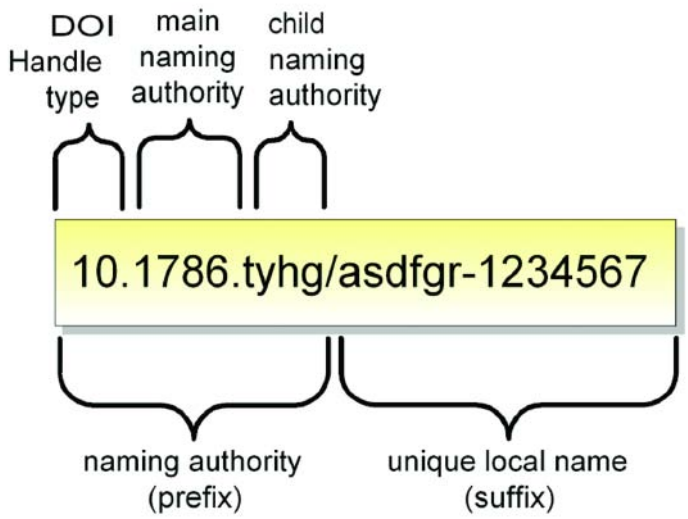
The handle architecture (which is further described in Appendix I) includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store names, or handles, of digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources.

Any of these associated values can be changed, as needed, to reflect the current state of the identified resource without changing the handle, thus allowing the name of the item to persist over changes of location and other current-state information.

Each handle may have its own administrator(s) and administration can be done in a distributed environment. The name-to-value bindings may also be secured, allowing handles to be used in trust management applications.

Every handle registered in LHS/GHS consists of two parts: its naming authority, otherwise known as its prefix, and its unique local name, otherwise known as its suffix. These parts are separated by a '/' (see Figure 8).

Figure 8: Handle Syntax



The naming authority identifies the administrative unit of creation.

Each naming authority may have many child naming authorities registered underneath it. Any child naming authority can only be registered by its parent at the global level – federated naming authority management is not yet possible – see section below.

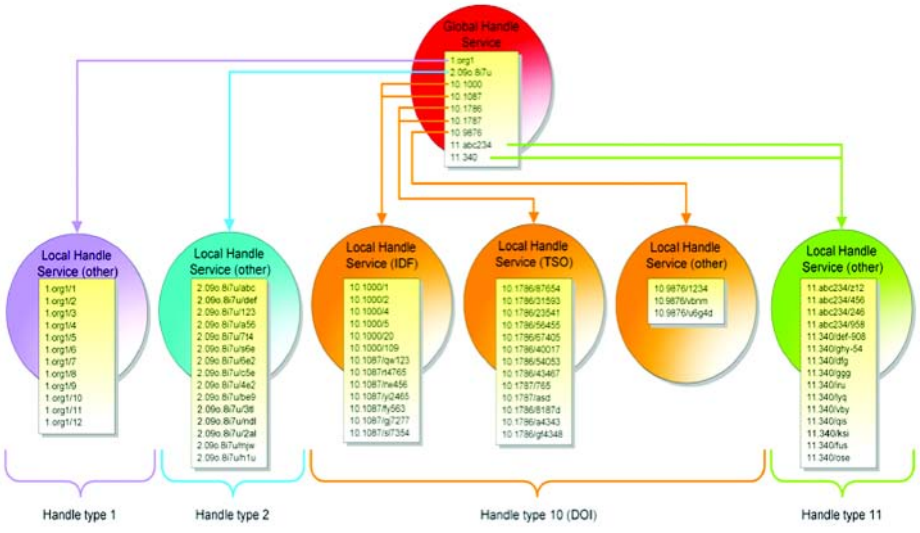
Every handle is defined under a naming authority. The uniqueness of a naming authority and a local name under that authority ensures that every handle is globally unique within the entire Handle System architecture.

The Handle System architecture is composed of two conceptual layers of look-up – a single Global Handle Service (GHS), and an unlimited number of Local Handle Services (LHSs) that it references. The GHS and LHSs are implemented using a physical architecture that removes single point of failures and allows limitless scalability.

The GHS is essentially a network aware, centralised look-up table that contains all handle prefixes, and the locations of all the corresponding LHSs, that collectively hold all the full handles (i.e. the prefixes and suffixes). The GHS and LHS repository structure and code base are identical – only their content differs as already stated.

A conceptual diagram of the GHS and LHSs can be seen in Figure 9. For clarity, each service is represented as a list of the handle records that it might contain, with appropriate connecting links.

Figure 9: Conceptual Handle System Architecture



As can be seen from Figure 9, the GHS as used by the DOI system is the same as used by all other handle implementations – be they public or not. Any handle is resolvable through the CNRI handle software, as long as its owner allows it to be. Thus, for example, some DTIC handles (not DOIs) can be resolved using the resolver at TSO. An example of this is the Handle 100.2/ADA300359, via the LHS handle resolver on the home page of www.tsoid.com.

The default LHS code distribution offered for prototyping uses a proprietary database format (developed by CNRI) that is optimised for look-up and does not support database administration. Once the total repository size starts to exceed 500,000 the replacement of the underlying handle database is straightforward scaling to enterprise levels.

Benefits of replacing the default handle database are:

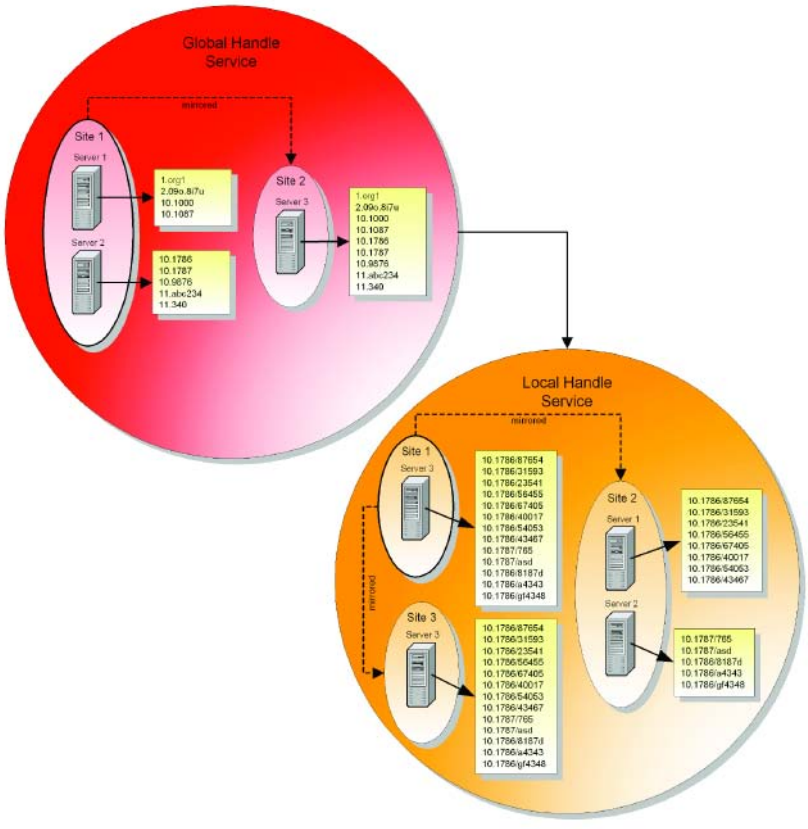
- Handles can be administered and queried. For example wildcard searching, mass handle inserts etc.
- Ability to use other database utilities such as backup, clustering, mirroring etc.
- Allows creation of a metadata repository with the same technology and is thus simpler to develop and maintain.

The Handle System is based on three layers of structure:

- SERVICE: the GHS and numerous LHSs Site/s
- SITE: A conceptual middle layer, describing one or more collections of actual machines
- SERVER: Physical layer of hardware used.

A representative example of the Service, Site and Server physical handle architecture can be seen in Figure 10. For clarity only one LHS is described.

Figure 10: Technical Handle System Architecture of a Notational Local Handle Service

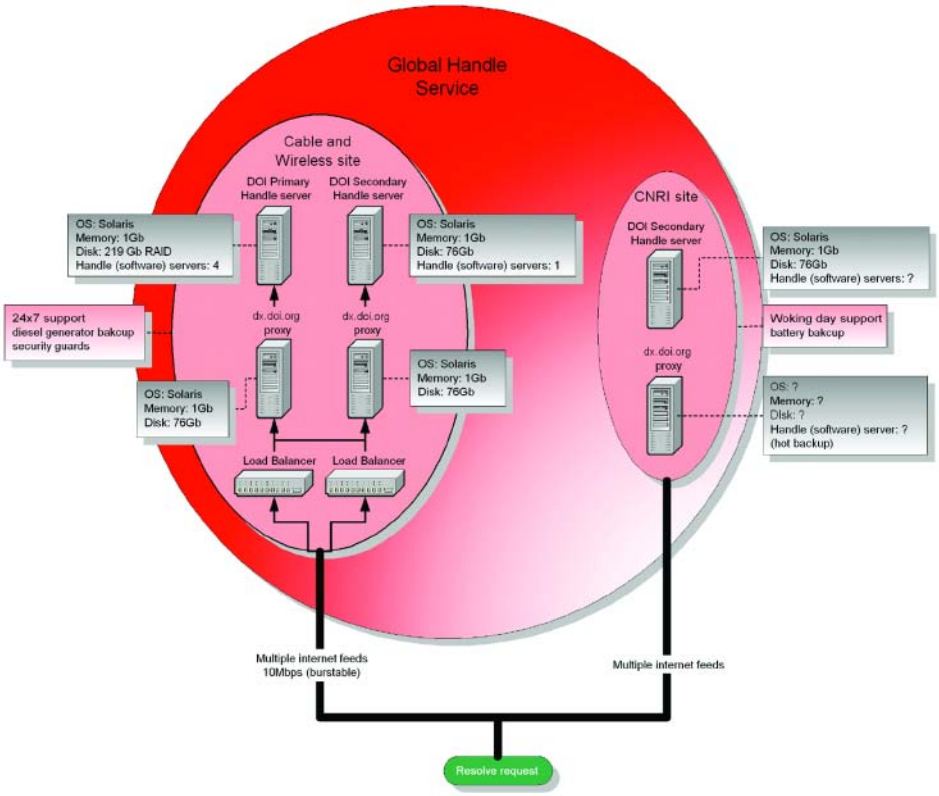


A hashing algorithm determines which physical machine (server) holds any given handle, and is determined by a parameter that has three possible values:

- Hash using the naming authority (prefix)
- Hash using the local name (suffix)
- Hash using the entire handle – i.e. prefix and suffix.

The hashing algorithm is based on the MD5 function, which is extensively used in cryptography. MD5 takes a variable-length character string (the handle) and produces a fixed-length hash, the IP address of the machine where the specific handle resides.

Figure 11: The Current GHS Physical Architecture



The current hardware and software employed in the GHS can be seen in Figure 11.

All handle code, including utilities and tools is in the region of 81,000 lines.

A.2 Delegated Handle System

The GHS/LHS two-layer implementation is to be modified to enable tree based, delegated (federated) LHS hierarchies. The data structures to allow this have always been present in the Handle System, but this capability has remained unimplemented as yet.

Federated LHSs will enable any LHS to create, independent of the GHS, its own naming sub-authorities and subsequent authentication process. Hence, the TSO LHS could create fully fledged naming authorities – for example 10.1786.departmentx, without the need to register and administrate this via the CNRI GHS, as is currently the case.

In general this is a good thing, as it devolves authority (literally) to any LHS at any level in a tree of LHS. It also makes the Handle System closer to the DNS system in this regard. However, one of the technical arguments for the Handle System over DNS has always been – 'Only two steps to every Handle'. This of course would not always be the case for federated LHS (caching aside) – the number of steps to resolve a handle would vary. Part of the recent National Science Foundation grant to CNRI is to develop this federated implementation.

Handle System Mirroring

Each handle service, including the Global Handle Registry (GHR), consists of at least one primary site and one or more secondary sites. All administration is done on the primary site and each secondary site is a mirror of the primary site. Each site, both primary and secondary, consists of one or more servers. The replication process does not require that each site has the same number of servers. The example given below illustrates the replication process using a service consisting of one primary and two secondary sites.

Figure 12

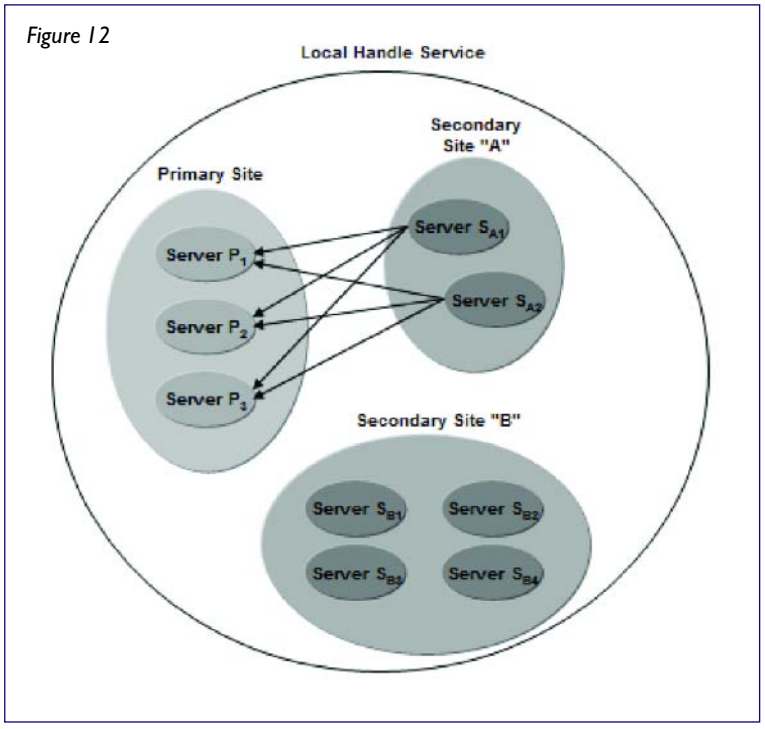
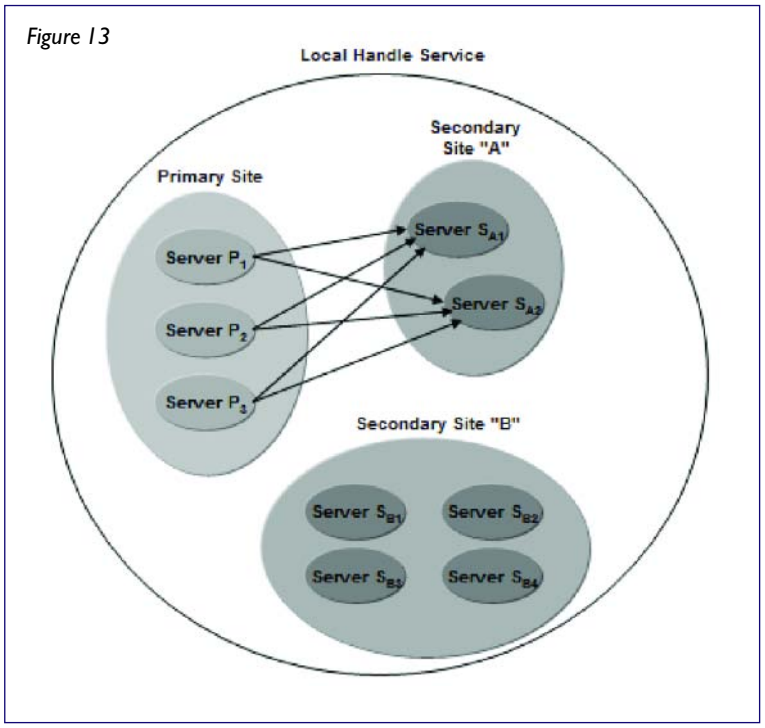


Figure 12 illustrates a local handle service with one primary site and two secondary sites. When secondary site 'A' began running, each server in the secondary site sent a request to each server in the primary site asking for updates.

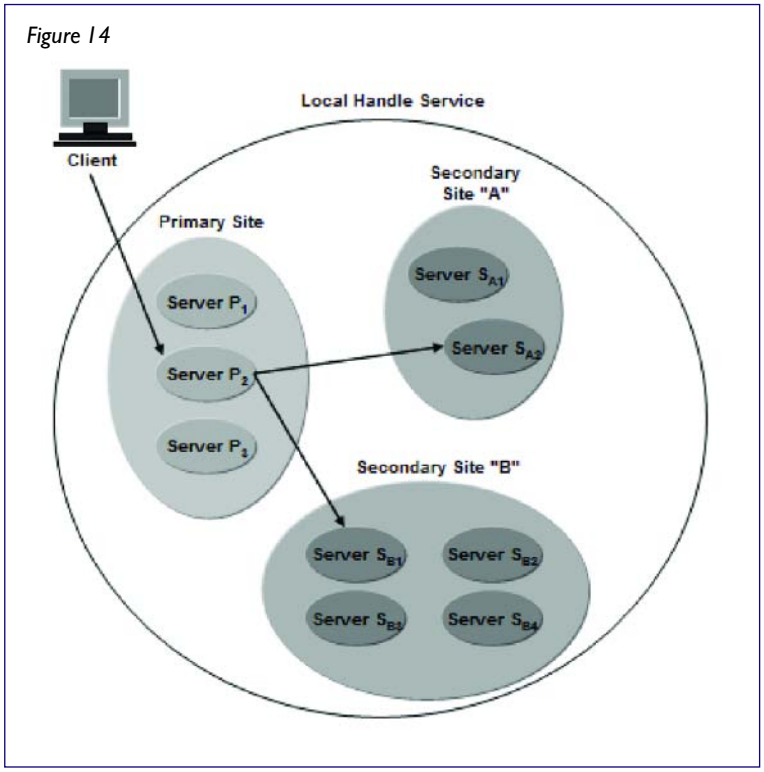
Figure 13



As shown in Figure 13 each server P_1 – P_3 knows which handles in its transaction log hash to which secondary server, and sends them. Each secondary will continue to request updates on a regular basis. The request is made in the form of 'all transactions since transaction X'. Thus the secondaries keep track of their update state and the mirroring process is mainly one of the secondaries pulling from the primaries.

In the example shown in Figure 14, for a given new administrative action, the admin client knows, because of hashing, that the action is performed on primary server P_2. Server P_2 then knows to send that action, when an update is requested, to secondary site 'A', server S_{A2} and secondary site 'B', server S_{B1}.

Figure 14


Local Handle Service
Client
Secondary Site "A"
Primary Site
Server S_{A1}
Server P_1
Server S_{A2}
Server P_2
Server P_3
Secondary Site "B"
Server S_{B1}
Server S_{B2}
Server S_{B3}
Server S_{B4}

14 Appendix B: Example XML Handles/DOI files

Example XML files for the download of DOIs, and upload of metadata from the TSO digital identifier site www.tsoid.com can be found below:

XML DOI download file example

```
<DOIExecutorService>
  <ReservedDOIs>
   <DOI>10.1790/571105700745</DOI>
   <DOI>10.1790/132263106124</DOI>
  </ReservedDOIs>
</DOIExecutorService>
```

XML metadata upload file example

```
<AssignDOIs>
  <MetadataRecord>
   <Identifier>10.1790/571105700745</Identifier>
   <Metadata>
     <Location>
      Cabinet XX, 3rd Floor, Stockley House,
      Wilton Road, Victoria, SW1V 1LQ
     </Location>
     <Title>e-GIF 1.0</Title>
     <Subject>e-Government</Subject>
     <Creator>Cabinet Office, UK</Creator>
     <Date>2000-10-01</Date>
     <PublicSearch>1</PublicSearch>
   </Metadata>
  </MetadataRecord>
  <MetadataRecord>
   <Identifier>10.1790/132263106124</Identifier>
   <Metadata>
    <Location>
      http://www.govtalk.gov.uk/schemasstandards/
      egif_document.asp?docnum=201
    </Location>
    <Title>e-GIF 2.0</Title>
    <Subject>e-Government</Subject>
    <Creator>Cabinet Office, UK</Creator>
    <Date>2001-04-05</Date>
    <PublicSearch>1</PublicSearch>
   </Metadata>
  </MetadataRecord>
 </AssignDOIs>
```

If you would like to find out more about any aspect of TSO
or its Information Management Services, contact:
The Stationery Office Ltd 51 Nine Elms Lane, London SW8 5DR
+44 (0) 870 600 5522 e-mail tsoservices@tso.co.uk

The Information Management Company

www.tso.co.uk