



Identificadores permanentes para o património cultural

É do conhecimento geral que os recursos da Internet têm uma vida curta; é a sua identificação e a sua complexa localização que levanta muitas questões tecnológicas e organizacionais envolvendo a citação, a recuperação e a preservação de recursos culturais/científicos. Isto não é só e apenas um problema técnico: a identificação de objectos digitais permanentes, incluindo textos, música, vídeo, imagens, documentos científicos e similares, é ainda mais problemática quando evita a utilização da actual internet como uma plataforma fiável para pesquisa e disseminação de conteúdos científicos e culturais.

Para que precisamos de “identificadores permanentes”?

A preservação a longo prazo, a disseminação e o acesso aos objectos culturais digitais, está agora entre as preocupações das instituições culturais como as universidades, os arquivos, os museus e as bibliotecas. A utilização de URL's não pode ser considerada uma abordagem fiável, para endereçar estes assuntos, devido á instabilidade estrutural dos links (por exemplo: quando os domínios deixam de estar disponíveis), bem como os recursos relacionados (devido a re-colocações ou a actualizações). A actual utilização da abordagem URL aumenta o risco de perda dos documentos culturais ou sub-utilizand colecções culturais disponíveis. No domínio da Herança Cultural (HC) é essencial não apenas identificar o recurso mas também garantir o seu acesso contínuo.

Uma solução fiável passa por associar um identificador permanente (IP) a um recurso digital que permaneça o mesmo, independentemente de onde este recurso esteja localizado.

Para implementar um sistema IP há alguns passos essenciais a concretizar:

- 1) Seleccionar os recursos que precisam de IP
- 2) Registrar os recursos identificados e registar a sua criação
- 3) Resolução do IP com o URL associado
- 4) Manutenção do registo que associa o IP – URL e garante um acesso continuo aos recursos

O primeiro passo é a prerrogativa de cada instituição cultural, considerando que os passos seguintes podem ser delegados a outras entidades de forma a assegurar uma mais eficaz sustentabilidade económica e funcional do serviço.

Requisitos de um sistema PI

Uma instituição HC deverá optar por uma infraestrutura que utilize no sistema os seguintes requisitos como directrizes:

- Singularidade global
- Persistência
- Resolutividade
- Fiabilidade
- Autoridade
- Flexibilidade
- Interoperabilidade
- Custos

Glossário

“Objecto” qualquer entrada de interesse numa transacção de propriedade intelectual é definida pelos metadados e pela terminologia de um dicionário de dados (index framework e tal). Os objectos podem ser físicos, digitais ou abstractos, por exemplo, pessoas, organizações, acordos, etc....

“Serviço de resolução” (de referência): o processo no qual um identificador está ligado é a ligação (a pedido) a um serviço de rede que receba em troca um resultado específico (recurso, metadados, etc. ...)

“Autoridade designada” é a autoridade independente que atribui os nomes, e garante a sua singularidade e persistência. Uma resolução de serviço de nomeação corresponde a todas as autoridades de nomeação e subentende o nome de resolução. Um sistema distribuído de IP's prevê que a responsabilidade de gerar e de resolver pode ser delegada noutras instituições denominadas autoridades sub-nomeadas que gerem uma parte do domínio/ espaço.

“Namespace” um contendor abstracto que providencia contextos para os itens que detém e permite a desambiguação de itens com o mesmo nome (e localizados em diferentes domínios).

“Registo” associação de nome entre URN's e um ou mais URL

“Repositório” local onde os recursos digitais são colocados (DSpace, Fedora, Codex, etc. ...) com ou sem um sistema de gestão de recursos (sistema de ficheiros)

“URI”, Uniform Resource Identifier Identificador de Recursos Uniforme é o conjunto genérico de todos os nomes/endereços que são curtas sequências no que se refere aos recursos.

“URL” Uniform Resource Locator Localizador de Recursos Universal é um URI que além de identificar um recurso, providencia meios de acção mediante a representação de um recurso pela descrição do mecanismo do seu acesso primário ou pela sua localização na rede.

Singularidade global

Consideramos o identificador como uma etiqueta que está associada ao objecto num determinando contexto. O “contexto” destina-se tanto ao tipo de padrão utilizado para o nome de sintaxe (por exemplo: URN:NBN:IT:xxx-xxxx) como para a identificação da autoridade que atribui esta etiqueta.

Persistência

Persistencia refere-se ao periodo de vida permanente de um identificador. Não é possível reatribuir o IP a outro recurso ou apagá-lo. Ou seja, o IP será globalmente único para sempre, e poderá ser utilizado como referência do recurso muito além do tempo de vida do recurso identificado ou da autoridade nomeada envolvida. A persistência é evidentemente uma questão específica do serviço ou da política de uma instituição cultural. A única garantia da utilidade e da persistência do sistema de identificador é o empenho demonstrado pela organização que atribui, gere e decide sobre os identificadores.

Resolubilidade

Resolubilidade refere-se à possibilidade de recuperar um recurso só este se estiver publicado. É importante distinguir o conceito de identificação do de resolução. A opção de identificar o resumo que contextualiza os itens não implica necessariamente que se escolha uma resolução arquitectural correspondente.

Fiabilidade

Para assegurar a fiabilidade de um sistema IP, devem ser avaliados dois aspectos: a infraestrutura IP deverá estar sempre activa (a redundância de serviços, os serviços do depósito das cópias de segurança, etc.), e o registo actualizado (através de sistemas automáticos).

Autoridade

A única garantia da utilidade e da persistência do sistema de identificador é o empenho demonstrado pela organização que atribui, gere e decide sobre os identificadores. No domínio HC a tendência é fazer uso da prestação de serviços das instituições públicas como bibliotecas nacionais, arquivos estatais, etc. Requisitos como a autoridade e a credibilidade de um sistema IP deverão ser cuidadosamente avaliados antes de adoptar uma solução.

Flexibilidade

Um sistema identificador será tão mais efectivo quanto mais for capaz de acomodar os requisitos especiais dos diferentes tipos de materiais ou de colecções. Por exemplo, um sistema identificador deverá ser capaz de gerir os diferentes níveis de granulosidade uma vez que um identificador deve apontar para os diferentes campos da aplicação do utilizador.

Interoperabilidade

Este aspecto é fundamental para garantir a possibilidade de difusão e de acesso aos objectos culturais digitais.

Muitas tecnologias e abordagens estão disponíveis e algumas delas são adaptadas aos requisitos de sectores específicos. Entre os diferentes sistemas, a interoperabilidade deve ser tida em conta, pelo menos ao nível dos serviços prestados que oferecem interfaces comuns e fáceis ao utilizador. O sistema de interoperabilidade pode basear-se na adopção de padrões abertos.

Tecnologias Correntes

PURL's (URL persistente) – Localizador de Recursos Uniforme Persistente. Funcionalmente, um PURL é um URL. Contudo, em vez de apontar directamente à localização de um recurso de Internet, um PURL aponta para um serviço de resolução intermédio através do redireccionamento das capacidades padrão do servidor Web que poderá redireccionar os pedidos de recursos utilizando o identificador persistente para a localização actual do documento ou do recurso.

www.purl.org

URN (Uniform Resource Name) – Identificadores Unívocos: UNR é a URI que utiliza o esquema URN e não implica a viabilidade do recurso identificado. Os URN's destinam-se a servir como identificadores de recursos persistentes e de localização independentes e são concebidos para tornar mais fácil para mapear outros namespaces (que partilham as propriedades dos URN's) no espaço URN. Portanto, a sintaxe URN fornece meios para codificar caracteres de dados num formulário que pode ser enviado em protocolos existentes, transcritos na maioria dos teclados, etc. ...

www.ietf.org/rfc/rfc11737.txt

“Handle System”: o Handle System é o nome de um serviço objectivo geral global que permite garantir a resolução do nome e da administração através de redes como a Internet. O Handle System gere meios, que são nomes únicos para os objectos digitais e outros recursos da InterNet. A autoridade nomeada é autorizada a criar e a manter instrumentos, o seu identificador deve ser único para essa autoridade mas não tem sintaxe prescritas.

www.handle.net

XRI (OAIS Extensible Resource Identifier): Identificador de Recurso Extensível – o objectivo do XRI é definir um esquema URI e o correspondente namespace URN para os serviços de distribuição do directório que permite a identificação dos recursos (incluindo pessoas e organizações) e a partilha de dados em vários domínios, empresas e aplicações.

www.oais-open.org

ARK (Archival Resource Key) (IETF Internet Draft): o esquema que pretende facilitar a nomeação persistente e a recuperação dos objectos de informação. O princípio base do ARK é que a persistência é apenas uma questão de serviço nem inerentes a um objecto nem conferidas a ele por uma simples nomeação de sintaxe. Identificador ARK resolve 3 diferentes saídas: recursos, metadados e compromisso de preservação.

www.cdlib.org/inside/diglib/ark

N2T (Name to thing) N2T é um consórcio de organizações de memória cultural e um pequeno e simples servidor Web espelhado em vários casos de fiabilidade a nível mundial. Este projecto pretende proteger 200 URL's de organizações da instabilidade do servidor com 200 regras reescritas pelos simples redireccionamentos HTTP para cada organização.

www.n2t.info

Custos

No domínio HC o sistema IP adoptado deverá ser livre de custos ou, pelo menos, com custos sustentáveis, uma vez que o papel das instituições culturais é o de garante do livre-acesso a recursos para além do tempo e evitar a fractura digital.

Outras Considerações

Granulosidade

A granulosidade refere-se ao nível de detalhe a que cada identificador permanente precisa de ser atribuído. Este requisito terá um impacto considerável no sistema de identificadores que a instituição adopte.

Nalguns casos, poderá ser necessário para citar uma página Web que serve de acesso a uma colecção de ficheiros Web, ou uma citação de um artigo de jornal, um item, ou um capítulo. Contudo, devido aos direitos de gestão podem ser exigidos alguns detalhes importantes. Cada instituição deverá avaliar se o serviço IP fornece o nível certo de granulosidade para o seu tipo de recursos.

PI's opacos ou semânticos

Um identificador permanente pode não conter informação o objecto que identifica (ID opaco) e isto porque é composto de caracteres aleatórios que não têm semântica associada. Um identificador opaco implica um trabalho de resolução para identificar, contudo pode conter um significado “construído” (ID semântico).

Genericamente é mais fácil memorizar e utilizar identificadores com uma base mnemónica do que aqueles que contenham uma sequência de caracteres sem sentido, embora isto não tenham qualquer relevância para o processamento da máquina.

Versões

Cada nova versão de uma fonte exige um identificador persistente em separado. Uma nova versão pode ser considerada como um objecto digital diferente porque o seu conteúdo ou formato físico pode ter sido alterado. Gerir diferentes versões pode ser possível através da definição de regras ou campos de metadados.

De que forma podem as tecnologias auxiliar-nos?

As aplicações IP exigem uma base de dados que registe o percurso da actual localização do objecto digital, denominado resolver database . Estes mapas localizam os recursos e redireccionam o utilizador à actual localização. O resolver database , bem como o seu serviço de resolução, poderão ser implementados sob diferentes formas: centralizado ou distribuído, baseado ou não no DNS¹.

Centralizado: esta arquitectura baseia-se num ponto central que gera o nome do recurso e assegura a sua resolução e confiança através do tempo. Esta solução implica a centralização de responsabilidades e gestão de custos; portanto, os serviços de resolução centralizada têm um único ponto de falha.

¹ Domain Name System – Sistema de Nomes de Domínios, é o sistema de gestão de nomes hierárquico, que opera mediante duas operações: examina e actualiza as bases de dados, ou define os nomes dos servidores em endereços de rede (IP's).

Resources

ERPANET workshop Persistent Identifiers

Thursday 17th - Friday 18th June 2004-University College
Cork, Cork, Ireland

www.erpanet.org/events/2004/cork/index.php

DCC Workshop on Persistent Identifiers

30 June – 1 July 2005

Wolfson Medical Building, University of Glasgow

<http://www.dcc.ac.uk/events/pi-2005/>

URN:NBN

<http://www.ietf.org/rfc/rfc3188.txt>

URN:NBN:DE

<http://www.persistent-identifier.de>

URN:NBN:IT

<http://www.rinascimento-digitale.it>

DOI

<http://www.doi.org>

ARK

<http://www.cdlib.org/inside/diglib/ark/>

PADI

<http://www.nla.gov.au/padi/topics/36.html>

PILIN

<https://www.pilin.net.au/>

OpenURL

http://www.niso.org/committees/committee_ax.html

Distribuído: esta arquitectura exige registos e serviços de resolução para cada autoridade sub-denominada nomeada para gerir os seus próprios nomes de IP's; a gestão de “nível máximo de autoridade”, redirecciona do processo para o serviço de resolução apropriado.

Baseado no DNS: o protocolo HTTP é utilizado para activar a citação ao link na rede através do requisito HTTP num serviço de resolução. Esta abordagem na base DNS não especifica clientes ou ligações para a normalização dos motores de busca.

Não baseado no DNS: mais implementações ajudaram a desenvolver um protocolo específico para a gestão de nomes e de resoluções de IP's (por exemplo: DOI). Neste caso, um cliente específico (ou uma ligação a um motor de busca), exige um identificador específico e o acesso ao objecto digital ou aos metadados associados. Esta solução permite uma aproximação para ampliar o serviço ao protocolo HTTP.

Oportunidades de Investigação

Com o crescimento das empresas de tecnologias de Informação, é dada cada vez mais atenção à questão da estabilidade do URL no que se refere aos acessos dos recursos na Internet. Os sistemas de identificadores persistentes são uma resposta relativamente recente a este problema. O contexto extremamente dinâmico onde estes sistemas operam, permite novos horizontes de investigação. Aqui ficam alguns aspectos interessantes e actualmente ainda por estudar em profundidade:

- a actual tendência é a adopção de sistemas que se relacionam com o uso domínio (por exemplo, NBN no domínio das bibliotecas). Contudo, o recurso pode ser parte de mais do que um domínio e pode ser identificado por sistemas diferentes. Deste modo é necessário garantir a interoperabilidade entre os diferentes sistemas de identificação e de implementações sedeadas no mesmo namespace ;

- Os Identificadores Persistentes permitem o acesso aos recursos mas também aos seus metadados, que são fundamentais para que o utilizador identifique o conteúdo. Portanto, é cada vez mais importante desenvolver uma gestão avançada de metadados e nos utilizadores dos serviços, como uma pesquisa extensível a diferentes repositórios.

- As relações semânticas entre os objectos multimédia podem ser tidos em consideração de forma a definir as ontologias e um melhor entendimento dos recursos da Internet.