MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

iconet.com.br/banon/2009/09.09.22.01

IDENTIFICADOR COM BASE NA INTERNET - PROJETO DE NORMA ABNT

Gerald Jean Francis Banon

URL do documento original:
<http://urlib.net/LK47B6W/362SFKH>

INPE
São José dos Campos

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

iconet.com.br/banon/2009/09.09.22.01

IDENTIFICADOR COM BASE NA INTERNET - PROJETO DE NORMA ABNT

Gerald Jean Francis Banon

URL do documento original:
<http://urlib.net/LK47B6W/362SFKH>

INPE
São José dos Campos

Dados Internacionais de Catalogação na Publicação (CIP)

Banon, Gerald Jean Francis.
B227 Identificador com base na Internet - Projeto de Norma ABNT / Gerald Jean Francis Banon. – São José dos Campos : INPE, .
40 p. ; (iconet.com.br/banon/2009/09.09.22.01)

1. Auditoria. 2. Memória científica. 3. Repositório digital. 4. Arquivo digital. 5. Biblioteca digital. I. Título.

CDU 021.61:657.6

AGRADECIMENTOS

Agradecemos aqui o Serviço de Informação e Documentação (SID) do INPE, e em particular à Marciana Leite Ribeiro, Silvia Castro Marcelino, Simone Angélica Del-Ducca Barbedo, Viveca Sant'Ana Lemos e Yolanda Ribeiro da Silva Souza, pelas sugestões de melhoria à primeira versão deste texto.

RESUMO

Esta Norma apresenta as várias formas em que um identificador global pode vir a ser utilizado para identificar e prover acesso consistente e perene a diversos tipos de itens de informação (documentos, mapas, imagens, etc.) armazenados em acervos, sejam eles repositórios ou arquivos digitais. A implantação e a utilização deste identificador global requerem, de forma direta, a infra-estrutura já existente na Internet, portanto, sem custo adicional, neste aspecto. O identificador global pode ser utilizado em associação com o processo de armazenamento de informação em acervos. O que torna simples a criação de cópias em acervos distintos e, também, a própria migração de itens de informação entre eles. As diversas aplicações de um identificador global desta natureza são de particular interesse em sistemas de dados espaciais e de informação.

ABSTRACT

This norm...,

SUMÁRIO

Pág.

LISTA DE TABELAS

Pág.

1 Introdução

1.1 Objetivo

Esta Norma estabelece as regras de construção de um identificador com base na Internet, assim como suas várias formas de apresentação.

1.2 Justificativa

Os hipervínculos (*hyperlinks*) ou simpelsmente vínculos ou ponteiros, elementos essenciais na navegação entre itens de informação (documentos, mapas, imagens, etc.) disponíveis na Internet devem ter seu funcionamento preservado por longo prazo.

A solução para tornar os ponteiros persistentes encontra-se no uso de um sistema de identificação global.

O sistema de endereçamento físico de um item de informação na Web por meio de uma URL (Uniform Resource Locator), não é um sistema de identificação, pois, com o tempo, um determinado item de informação pode mudar de localização, fazendo com que a associação item de informação $\mapsto$ URL não fique permanente.

Uma vez escolhido um sistema de identificação e por meio dele atribuido identificadores à itens de informação, o problema da construção de ponteiros persistentes pode ser solucionado por meio do uso de um resolvedor de identificação, cujo papel consiste em redicionar cada URL, agora contendo apenas o identificador de um item de informação, para a URL contendo o seu endereço físico.

O sistema de identificação descrito nesta Norma apresenta-se como uma alternativa simples, quando comparada a outras soluções como, por exemplo, o PURL ou o DOI®. Ele está sendo extensivamente utilizado, desde 1995, na plataforma UR*Lib*.

2 Terminologia

Identificador de um item: o **rótulo** atribuido à um **item** por um **sistema de identificação**.

Item: qualquer objeto a ser identificado.

Item de informação: qualquer **item** contendo dados, isto é quaisquer dados a serem identificados. Por exemplo: documentos, mapas, imagens, etc.

Repositório uniforme de um item: **Identificador de um item** usado para armazená-lo.

Rótulo: qualquer sequência finita de caracteres escolhidos dentro de um alfabeto finito, usada na identificação de um **item**.

Sistema de identificação: qualquer mapeamento injetor entre um conjunto de **itens** e um conjunto de **rótulos** (ver detalhes na Seção 3).

Sistema de identificação em dois níveis: qualquer **sistema de identificação** associando um **item** à um **rótulo** obtido a partir de um par de **rótulos**, o primeiro identificando o **subsistema de identificação** responsável pela identificação do **item**, e o segundo sendo o **rótulo** atribuido ao **item** por esse **subsistema de identificação** (ver detalhes na Seção 3).

Subsistema de identificação: qualquer **sistema de identificação** restrito a um subconjunto de **itens**.

3 Construção de um sistema de identificação

Nessa Norma, os **itens** (objetos a serem identificados) são considerados formando conjuntos. Por exemplo, conjunto de computadores possuindo um IP (*Internet Protocol*) fixo, conjunto de pastas, etc. Por sua vez, os **rótulos** identificando os **itens**, são considerados formando conjuntos finitos ou enumeráveis. Por exemplo, o conjunto das sequências de no máximo 255 caracteres alfanuméricos, ou ainda ou conjunto dos números decimais representando os anos.

Pela restrição dos conjuntos de **rótulos** serem finitos (resp. enumeráveis), e pela propriedade do **sistema de identificação** ser injetor, os conjuntos dos **itens** devem ser necessariamente finitos (resp. enumeráveis).

Por ser um mapeamento injetor, um **sistema de identificação** associa, de forma permanente, cada **item** à um único **rótulo**, de maneira que, **itens** distintos sejam associados à **rótulos** distintos.

Considerando o **sistema de identificação** como um processo dinámico (i.e., que ocorre ao longo do tempo), a solução geral para montar um mapeamento injetor consiste em atribuir um novo **rótulo** à cada novo **item**.

Uma solução particular consiste em utilizar como **identificador de um item**, a data e hora na qual é feita a atribuição, sendo a granularidade da data suficientemente fina para distinguir entre duas atribuições sucessivas.

No entanto, as soluções anteriores, pressupõem um **sistema de identificação** centralizado, sob a responsabilidade de um único ator, para o qual toda requisição de nova identificação deve ser encaminhada, tornando o **sistema de identificação** relativamente vulnerável.

Por este motivo, uma solução segura deve levar à um **sistema de identificação em dois níveis**. O primeiro nível consiste em um único **sistema de identificação** enquanto o segundo nível consiste em vários **subsistemas de identificação** cada um sob a responsabilidade de um ator distinto.

Num primeiro momento, cada **subsistema de identificação** (visto como objeto a ser identificado) recebe, do **sistema de identificação** do primeiro nível, um **rótulo**. Num segundo momento, cada **item** (do conjunto de itens de interesse) recebe,

do **subsistemas de identificação** responsável por esse **item**, um **rótulo**. A concatenação desses dois **rótulos** constitue o **identificador do item** (ver Proposição X em anexo).

Enquanto o **identificador de um item** fornecido por um determinado **subsistemas de identificação** tem validade apenas dentre do escopo deste subsistema, o **identificador desse item** obtido por concatenação, tem validade dentre do escopo global de todos **subsistemas de identificação**.

O Handle System®, por exemplo, funciona desta maneira.

Usando a terminologia deste sistema, a identificação final é a concatenação de um "prefixo" identificando um **subsistema de identificação** e de um "sufixo" identificando um **item** no escopo deste **subsistema de identificação**.

A solução apresentada aqui, consiste em reaproveitar a infra-estrutura já existente da Internet para identificar os **subsistemas de identificação**, portanto, sem custo adicional, neste aspecto.

Supondo que cada **subsistema de identificação** seja hospedado em um computador ligado à Internet com IP fixo, ele é naturalmente identificado pelo seu endereço nesta rede, fornecendo assim, de forma simples, o prefixo.

Afim de desburocratizar ainda mais o **sistema de identificação** como um todo, o sufixo, fornecido por um **subsistema de identificação**, segue uma regra comum a todos os subsistemas, e consiste na data e hora da associação do item de informação ao sufixo. Esta escolha facilita o reuso de um identificador de **subsistema de identificação** quando este passa sob o controle de um novo ator.

Na solução objeto desta Norma, dois tipos de prefixo herdado da Internet são considerados.

O primeiro tipo consiste em adotar como **identificador de um subsistemade identificação**, isto é como prefixo, o nome de domínio do computador que o hospeda, assim como sua porta de acesso.

Na Seção 3.1 entitulada *Name space specifications and terminology*, Mockapetris (1987) define o conceito de "nome de domínio" (*domain name*), e na Seção 3.2.2 entitulada *Server-based Naming Authority*, Berners-Lee et al. (1998) definem o con-

ceito particular de "nome de domínio de um computador" (*hostname*).

No segundo tipo, o prefixo é obtido utilizando o IP do computador, no lugar do nome de domínio.

Os exemplos reais a seguir antecipam alguns detalhes sobre a formação dos identificadores que serão dados nos dois próximos capítulos.

Exemplo 1 (identificador com base no nome de domínio).

A associação do *item* ao sufixo, ocorrida em 14 de abril de 2008 às 11 horas 53 minutos, resultou no sufixo:

`2008/04.14.11.53`

O **subsistema de identificação** emitindo este sufixo era hospedado em um computador com nome de domínio `md-m09.sid.inpe.br`, e acessível a partir da porta `80`, levando ao uso do prefixo:

`sid.inpe.br/md-m09@80`

Desta forma, o **identificador para o item** passou a ser:

`sid.inpe.br/md-m09@80/2008/04.14.11.53`

Usando, por exemplo, o resolvedor de identificação `urlib.net`, o ponteiro (URL) persistente para este *item* ficou:

`http://urlib.net/sid.inpe.br/md-m09@80/2008/04.14.11.53`

Adicionalemente, o ponteiro (URL) persistente para os metadados deste *item* ficou:

`http://urlib.net/sid.inpe.br/md-m09@80/2008/04.14.11.53??`

Observa-se, que mesmo que o nome de domínio `md-m09.sid.inpe.br` passe a ser abandonado ou muda de dono, isto não inviabiliza o **identificador criado para esse item**. O importante, apenas, é que estes dados eram pertinente no contexto da Internet na data e hora da associação entre o **item** e seu **rótulo**. Esta observação vale também para o segundo exemplo a seguir.

Exemplo 2 (identificador opaco com base no IP).

A associação do *item* com o sufixo, ocorrida em 16 de fevereiro de 2009 às 17 horas 46 minutos, resultou no sufixo opaco:

`34PGRBS`

O **subsistema de identificação** emitindo este sufixo era hospedado em um computador com IP `150.163.34.243`, e acessível a partir da porta `800`, levando ao uso do prefixo opaco:

`8JMKD3MGP8W`

Desta forma, o **identificador para o item** passou a ser:

`8JMKD3MGP8W/34PGRBS`

Usando, por exemplo, o resolvedor de identificação `urlib.net`, o ponteiro (URL) persistente para este *item* ficou:

`http://urlib.net/8JMKD3MGP8W/34PGRBS`

Adicionalemente, o ponteiro (URL) persistente para os metadados deste *item* ficou:

`http://urlib.net/8JMKD3MGP8W/34PGRBS??`

Observa-se que a granularidade do prefixo é extremamente fina já que os **subsistemas de identificação** são atrelados à números de porta de computador possuindo nome de domínio ou IP.

Quanto a granularidade do sufixo ela pode ser aumentada sem dificuldade, acrescentando por exemplo os segundos.

Os **sistemas de identificação** apresentados a seguir são agrupados em duas categorias. Na primeira, o **identificador do item** exibe o nome de domíno, e é chamado de **repositório uniforme do item**. Na segunda categoria, o **identificador do item** é construido com base no IP e tem a propriedade de ser opaco porque não exibe explicitamente nem o IP, nem o número de porta, nem a data e hora da criação do identificador.

4 Regras de construção do identificador como repositório uniforme

No sistema de identificação apresentado nesta seção, o identificador é chamado também de "repositório uniforme" porque ele pode ser usado para definir uma seqüência de quatro diretórios servindo para armazenar, num sistema de arquivos, o item de informação sendo identificado.

Os repositórios são chamados de uniforme porque eles são criados usando uma mesma regra de construção por todos os subsistemas. Desta forma, qualquer repositório pode ser instalado em qualquer subsistema, debaixo de um mesmo diretório, sem conflito de nome.

Num identificador como repositório uniforme, o prefixo e o sufixo são separados por `"/"` e cada um é, por sua vez, subdividida em duas partes separadas também por `"/"`. Assim, o identificador é constituida de quatro partes, formadas por, nesta ordem:

a) um nome de subdomínio,

b) um rótulo de domínio, e eventualemente um número de porta, separados por `"."` ou por `"@"`,

c) um ano e

d) um mês, dia, hora, minuto, e eventualmente segundo, separados por `"."`.

Essas quatro partes são reconhecíveis no Exemplo 1 do capítulo anterior, onde o identificador como repositório uniforme era:

```
sid.inpe.br/md-m09@80/2008/04.14.11.53
```

Para definir precisamente a sintaxe de um identificador como repositório uniforme, nessa norma, usa-se uma gramática BNF (aumentada) (CROCKER, 1982; CROCKER; OVERELL, 2008) com a seguinte alteração: `"|"` é utilizado para alternativas no lugar de `"/"`.

```
   repositório  =  prefixo "/" sufixo
                        ; ex: sid.inpe.br/mtc-m19/2010/08.25.12.38
       prefixo  =  subdomínio "/" rótulo [("." | "@") porta]
                        ; ex: sid.inpe.br/mtc-m19
    subdomínio  =  *(rótulo ".") último-rótulo ["."]; ex: dpi.inpe.br
        rótulo  =  ALFANUM | (ALFANUM *(ALFANUM | "-") ALFANUM); ex: sid
       ALFANUM  =  ALFA | DÍGITO
          ALFA  =  ALFAMI | ALFAMA
        ALFAMI  =  "a" | "b" | "c" | "d" | "e" | "f" | "g" | "h" | "i" |
                   "j" | "k" | "l" | "m" | "n" | "o" | "p" | "q" | "r" |
                   "s" | "t" | "u" | "v" | "w" | "x" | "y" | "z"
        ALFAMA  =  "A" | "B" | "C" | "D" | "E" | "F" | "G" | "H" | "I" |
                   "J" | "K" | "L" | "M" | "N" | "O" | "P" | "Q" | "R" |
                   "S" | "T" | "U" | "V" | "W" | "X" | "Y" | "Z"
        DÍGITO  =  "0" | "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" |
                   "9"
último-rótulo   =  ALFA | (ALFA *(ALFANUM | "-") ALFANUM); ex: br
         porta  =  1*DÍGITO; ex: 80
        sufixo  =  ano "/" mês "." dia "." hora "." minuto ["." segundo]
                        ; ex: 2010/08.25.12.38
           ano  =  4*DÍGITO; ex: 2010
           mês  =  2DÍGITO; ex: 08
           dia  =  2DÍGITO; ex: 25
          hora  =  2DÍGITO; ex: 12
        minuto  =  2DÍGITO; ex: 38
       segundo  =  2DÍGITO
```

A regra <subdomínio> é denotada <hostname> em Berners-Lee et al. (1998)

- a expressão <rótulo-de-domínio>.<nome-de-subdomínio> deve ser o nome de domínio do computador hospedando o subsistema de identificação, e o <número-de-porta> deve ser a porta de acesso a esse subsistema;

- o <ano> e o <mês>.<dia>.<hora>.<minuto>[.<segundo>], expressos numericamente, devem ser à data/hora GMT (*Greenwich Mean Time*) da criação do repositório pelo subsistema;

- ano = 4dígito

- mês = 2dígito

- dia = 2dígito

- hora = 2dígito

- minuto = 2dígito

- segundo = 2dígito

- dígito = 0|1|2|3|4|5|6|7|8|9

5 Comparação com o Handle System® e o DOI®

Como no Handle System®, o identificador usado na UR*Lib* possui um prefixo e um sufixo separado por uma barra "/". No entanto, a principal diferença reside no modo de geração do prefixo.

No sistema de identificação usado na UR*Lib*, o cadastramento dos provedores de dados junto ao resolvedor de identificação não é um pre-requisto para os provedores de dados começarem a trocar dados entre si, por meio de importação de cópias e inclusão de vínculos relativos.

Isto é possível porque no identificador usado na UR*Lib* o cadastramento dos prefixos é herdado do próprio funcionamento da Internet como meio de comunicação entre atores já previamente registrados. Com isto, a geração dos identificadores pode ser feita pelos próprios provedores de dados, sem necessidade de prévio cadastramento junto ao resolvedor de identificação.

Por exemplo, considerando um caso real, antes mesmo de se registrar junto ao resolvedor de identificação, o provedor com prefixo iconet.com.br/banon pôde importar, sem conflito de nomes, do provedor com prefixo dpi.inpe.br/banon uma cópia do documento identificado por dpi.inpe.br/banon/1998/08.02.08.56.

No provedor com prefixo iconet.com.br/banon, foi também possível incluir no arquivo: iconet.com.br/banon/2003/11.21.21.08/doc/cgi/oai.tcl, um vínculo relativo para o documento importado mencionado acima e contendo o arquivo doc/utilities1.tcl. Este vínculo apresenta-se da seguinte forma: ../../../../../../dpi.inpe.br/banon/1998/08.02.08.56/doc/utilities1.tcl.

Neste exemplo, observa-se que, usando o identificador global usado na UR*Lib*, foi possível, sem necessidade de recorrer ao sistema de resolução de nome, estabelecer um vínculo entre dois documentos originalmente depositados em dois provedores distintos.

O modo de geração do sufixo também é diferente. No caso do Handle System®, o modo de geração é livre e por conta do ator registrado neste sistema. No sistema de identificação usado na UR*Lib*, o modo de geração é padrão e o mesmo para todos os atores (ver Seção 3). Este último sistema é interessante, pois prevê inclusive os casos em que um novo ator se apropria de um prefixo caído em desuso. Naquele momento, este novo ator não precisa tomar conhecimento do sistema de geração dos identificadores utilizado pelo antigo ator, nem ter o cuidado de continuar a usá-lo, basta utilizar o sistema de geração padrão.

Finalmente, em comparação com o DOI® 3, observa-se que tanto o DOI® quanto o identificador usado na UR*Lib* possuem múltiplos níveis de resolução. No entanto, este último resolve também a distinção entre um documento e suas cópias, oferecendo para o usuário final a garantia que o documento acessado é sempre o mesmo, seja

ele uma cópia ou não.

REFERÊNCIAS BIBLIOGRÁFICAS

BERNERS-LEE, T.; FIELDING, R.; IRVINE, U. C.; MASINTER, L. **Uniform Resource Identifiers (URI): Generic syntax**. Washington DC: The Internet Engineering Task Force (IETF), Aug. 1998. 40 p. RFC 2396. Disponível em: <http://tools.ietf.org/html/rfc2396>. Acesso em: 19 ago. 2010. 6, 12

CROCKER, D. H. **Standard for the format of ARPA Internet messages**. Washington DC: The Internet Engineering Task Force (IETF), Aug. 1982. 47 p. RFC 822. Disponível em: <http://tools.ietf.org/html/rfc822>. Acesso em: 19 ago. 2010. 11

CROCKER, D. H.; OVERELL, P. **Augmented BNF for Syntax Specifications: ABNF**. Washington DC: The Internet Engineering Task Force (IETF), Jan. 2008. 16 p. RFC 5234. Disponível em: <http://tools.ietf.org/html/rfc5234>. Acesso em: 19 ago. 2010. 11

MOCKAPETRIS, P. **Domain names - concepts and facilities**. Washington DC: The Internet Engineering Task Force (IETF), Nov. 1987. 55 p. RFC 1034. Disponível em: <http://tools.ietf.org/html/rfc1034>. Acesso em: 19 ago. 2010. 6

APÊNDICE A - DEFINIÇÕES E PROPRIEDADES

Uma função

Definição 1 (definição de função)**.** Sejam X e Y dois conjuntos não vazios. Uma *função* f com *domínio* X e *contradomínio* Y associa cada elemento x de X à um único elemento y de Y. O elemento y é denotado por $f(x)$ e chamado de *valor de* f *em* x. Em outros termos, f satisfaz os seguintes axiomas:

(i) para todos x em X, existe um y em Y tal que $y = f(x)$;
(ii) para todos x_1 e x_2 em X, $f(x_1) \neq f(x_2) \Rightarrow x_1 \neq x_2$.

Uma função f de X em Y é denotada por $f : X \to Y$. □