CATCH Plus: Persistente Identifiers

Hennie Brugman
2009-06-03

Probleemstelling

Archiefsystemen en andere informatiesystemen bij erfgoedinstellingen bevatten vele objecten waarnaar verwezen moet kunnen worden. Deze objecten hebben daarom een naam of nummer: een *identifier*. Het kan hierbij zowel gaan om fysieke objecten als om digitale objecten. Catalogi bevatten metadata-beschrijven, die met behulp van een identifier verwijzen naar objecten. Metadata-beschrijvingen bevatten vaak verwijzingen naar concepten in een vocabulaire ("thesaurus-termen"). Deze concepten hebben ook een identifier. Verder kan het wenselijk zijn de metadata-beschrijvingen zelf ook van een identifier te voorzien. Andere objecten, die in de context van CATCH Plus projecten en diensten een identifier nodig hebben zijn annotaties en gebruikersprofielen.

In de praktijk zijn er binnen de erfgoedinstellingen vele verschillende identifier-systemen in gebruik. Meestal zijn zulke identifiers slechts *uniek* in de context van één specifiek informatiesysteem, soms is zelfs dat niet het geval. In een toenemend aantal gevallen zijn identifiers uniek binnen de context van een instelling. Verder veranderen identifiers van objecten nog al eens, bijvoorbeeld als gegevens worden ge-exporteerd of gemigreerd naar een nieuw database management systeem. Identifiers zijn dus vaak niet *persistent.* Dat is problematisch als van elders verwijzingen worden gemaakt met behulp van zo'n identifier. In de dagelijkse praktijk van data management en archiveren kost het oplossen van dergelijke problemen veel tijd en geld.

Het toenemend belang van internet en het web, en van samenwerkingsverbanden als CATCH Plus versterken het probleem: als er van buiten de instellingen naar je objecten wordt verwezen, hoe garandeer je dan als instelling dat de gebruikte identifiers uniek en persistent zijn? Als instellingen data en diensten in samenwerkingsverbanden of als consortium gaan ontwikkelen en aanbieden, hoe verwijs je dan vanuit je eigen instelling betrouwbaar en duurzaam naar die externe data, zodat je eigen bedrijfsprocessen of diensten niet in gevaar komen? Hoe om te gaan met objecten die verhuizen van instelling?

Het is in de eerste plaats noodzakelijk goede afspraken te maken over richtlijnen en verantwoordelijkheden met betrekking tot het toekennen en beheren van identifiers. In de tweede plaats kan dan de uitvoering van zulke afspraken zo eenvoudig, betrouwbaar en efficiënt mogelijk worden gemaakt met behulp van ondersteunende technologie.

Dit document beoogt een eerste aanzet te zijn tot het in CATCH Plus verband formuleren van afspraken en richtlijnen, en stelt bijpassende technologische oplossingen voor.

Algemene oplossingen

Introduceren en gebruiken van persistente identifiers is in de eerste plaats een organisatorische opgave en pas in de tweede plaats een technisch probleem. Iedere oplossing gaat gepaard met doorlopende administratieve taken en dus kosten. Zonder dat deze beheerstaken goed geregeld zijn gaat geen enkele technologische oplossing werken.

Beleidskeuzen

Het invoeren van persistente identifiers begint met het in onderling overleg maken van afspraken over een aantal punten. Deze afspraken moeten worden gedocumenteerd en voor alle betrokken inzichtelijk beschikbaar worden gesteld. Vragen die moeten worden beantwoord zijn onder meer:

Welke objecten moeten kunnen worden geassocieerd met een identifier?
Moet er wel of niet onderscheid worden gemaakt tussen fysieke objecten en hun gedigitaliseerde representaties? Moeten metadata records een eigen identifier hebben? Willen we samengestelde objecten kunnen identificeren? Queries? Gebruikers? Concepten en alignments uit de gezamenlijke vocabulaire repository? Moeten we naar segmenten van digitale objecten kunnen verwijzen? Naar annotaties? Omgekeerd, voor welke objecten is het niet nodig persistente identifiers te gebruiken?

Wat is het formaat van een identifier?
Welke lettertekens zijn toegestaan? Zijn ze ook bedoeld voor menselijke inspectie? Mogen ze betekenisvolle informatie bevatten? Zo ja, welke informatie wel en welke niet? Mag bijvoorbeeld de naam van de organisatie van herkomst erin opgenomen worden? Bestaan er al identifiers binnen deelnemende organisaties, die opgenomen kunnen/moeten worden? Gaan we identifiers automatisch genereren?

Welke personen en organisaties zijn verantwoordelijk voor het toekennen en beheren van identifiers?
Wie mogen nieuwe identifiers uitgeven? Wie mogen deze identifiers veranderen (met een nieuwe URL associëren)? Wie mogen dergelijke rechten delegeren? Wie gaat over het beheer van gebruikers en groepen? Het gaat hier typisch over taken die horen bij (digitaal) collectie-beheer.

Wanneer moet aan een object een nieuwe identifier worden toegekend?
Het gaat hierbij om beheer van versies en om identificeren van varianten van een digitaal object (bijvoorbeeld html, xml en pdf versies van hetzelfde object).

Hoe om te gaan met verhuizen van objecten of collecties tussen instellingen?
Krijgen die objecten nieuwe identifiers (bij voorkeur niet)? Welke procedures zijn er te volgen?

Persistentie
Persistentie heeft een aantal aspecten: ten eerste moeten identifiers zelf onveranderd blijven over de tijd. Ten tweede moeten identifiers resolvable blijven (bijbehorende locaties moeten bekend zijn). Ten derde moeten geïdentificeerde objecten gevonden kunnen worden op de locaties die door de resolver worden aangegeven. Tenslotte moet een identifier in de loop van de tijd hetzelfde object blijven identificeren. Voor al deze aspecten moeten door de deelnemende organisaties garanties worden afgegeven met betrekking tot persistentie.

Hosting
Gaan we zelf identifiers hosten, en zo ja, wie neemt de verantwoordelijkheid met betrekking tot beheer en beschikbaarheid van bijbehorende diensten? Gaat één instelling hosten, of meerdere? Zo nee, welke centrale autoriteit gaat het dan doen? Welke mate van controle willen we dan over onze eigen identifiers? Welke garanties kan zo'n autoriteit geven? Een alternatief is het periodiek *harvesten* van identifiers. (let op: hosting van identifiers is iets anders dan beheren van de identifiers)

Technologische ondersteuning

Er zijn momenteel een aantal min of meer wijdverbreide oplossingen beschikbaar om persistente identifiers technologisch te ondersteunen. De meest relevante en bekende zijn URN-NBN[1], Handles[2], DOI[3], PURL[4], en ARK[5].

Met uitzondering van URN-NBN kennen al deze oplossingen een bijbehorende *resolver*-architectuur (er bestaan resolver-oplossingen voor URN-NBN, maar die verschillen van land tot land). Het basis-idee achter een resolver is, dat alle identifiers voor een bepaalde 'Naming Authority' bij elkaar in één (virtuele) repository worden opgeslagen, die onder het beheer van de NA valt. Voor iedere identifier wordt extra informatie opgeslagen, minimaal de locatie(s) waar het bij een identifier behorende object kan worden gevonden. Zo'n locatie heeft typisch de vorm van een URL. Het grote voordeel van een dergelijke oplossing is dat identificatie en locatie van elkaar worden losgekoppeld: als de locatie(s) van een object wordt veranderd moet alleen de tabel in de resolver worden aangepast, alle verwijzingen naar het object gebruiken de identifier, en blijven dus automatisch geldig.

Diensten die zo'n systeem voor persistente identifiers kan vervullen, en andere mogelijke eigenschappen ervan:

1. het automatisch genereren en/of valideren van (wereldwijd) unieke identifiers
2. het 'resolven' van identifiers
3. met een bepaald domein kan expliciet een 'Naming Authority' worden geassocieerd. Verantwoordelijkheid voor toekennen en beheren van identifiers wordt expliciet neergelegd bij een instelling of consortium. De NA

[1] http://www.ietf.org/rfc/rfc3188.txt
[2] http://www.handle.net/
[3] http://www.doi.org/
[4] http://purl.org/
[5] http://www.cdlib.org/inside/diglib/ark/

    a. gaat commitments aan met betrekking tot het garanderen van persistentie van identifiers
    b. bepaalt zelf het beheer over PI-URL associaties
        i. Hoe zien PIs eruit?
        ii. Wat identificeren ze?
        iii. Wie heeft rechten om ze te veranderen?
4. Persistente identifiers kunnen worden geassocieerd met metadata (bijvoorbeeld de beheerder van de PI, een beschrijving, een email adres of een statement met betrekking tot garanties voor persistentie)
5. Het updaten van de locatie(s) die met een PI is geassocieerd door geauthoriseerde personen
6. fijnmazige oplossingen voor het organiseren van redundantie (bv door mirroring)
7. identifiers blijven bestaan, zelfs nadat geïdentificeerde objecten zijn verdwenen
8. de geschiedenis van een persistente identifier kan worden bijgehouden.

Eisen en wensen vanuit CATCH Plus

Hieronder worden een aantal requirements voor persistente identifier-oplossingen binnen de context van CATCH Plus genoemd en kort toegelicht.

Algemeen

Bij voorkeur aansluiten bij bestaande initiatieven voor uitgifte en beheer van persistente identifiers
Momenteel zijn twee lopende initiatieven in de omgeving van CATCH Plus bekend: nationale bibliotheken in Europa (waaronder de Koninklijke Bibliotheek) werken samen aan de invoering van persistente identifiers op basis van URN-NBN. Binnen de Nederlandse context (maar in overleg met andere Europese partijen) heeft DANS[6] een resolver-oplossing geïmplementeerd. Er is momenteel alleen commitment met betrekking tot het aanbieden van resolver-diensten binnen de context van SurfShare. Deze commitment kan momenteel nog niet worden waargemaakt i.v.m. ontbrekende (redundante) servercapaciteit.
Binnen Europa wordt momenteel gewerkt aan een zelfstandige en redundante identifier oplossing voor eScience op basis van Handles. Een tweetal organisaties heeft zich hieraan al gecommitteerd (de Duitse Max-Planck-Gesellschaft[7], via de GWDG[8], en het Finse CSC[9]). Er worden nog enkele partners gezocht, waaronder één in Nederland.

Gebruik maken van bestaande en beproefde technologische oplossing(en)
De meest gangbare oplossingen zijn momenteel URN-NBN, Handles, DOI, ARK en PURL.

Stem af op andere erfgoed-infrastructuurprojecten.
CATCH Plus dient bij te dragen aan de uiteindelijk digitale infrastructuur voor het Nederlandse erfgoed. Daarom moet met name voor essentiële infrastructurele aspecten samenwerking worden gezocht met andere projecten op het terrein van digitaal erfgoed. Met name gaat het hier om NED![10] en Europeana[11].

Bij voorkeur een homogene oplossing voor het CATCH Plus consortium
Momenteel is, voor zover bekend, alleen bij de Koninklijke Bibliotheek een oplossing voor persistente identifiers binnenshuis in gebruik. Dit biedt een goede kans om gezamenlijk een homogene oplossing in te voeren. Dit komt de interoperabiliteit ten goede, en maakt het delen van kennis en ervaring eenvoudig.

Inclusief met betrekking tot anderssoortige en al bestaande persistente identifiers
Instellingen hebben hun eigen eisen en wensen, en nemen ook deel aan andere samenwerkingsverbanden. De normale situatie zal dan waarschijnlijk uiteindelijk ook zijn dat meerdere oplossingen voor PIs naast elkaar bestaan. Een (homogene) oplossing, waarvoor in CATCH Plus gekozen gaat worden, moet dan ook compatibel zijn met andere PI systemen.

---

[6] http://www.dans.knaw.nl/
[7] http://www.mpg.de/
[8] http://www.gwdg.de
[9] http://www.csc.fi/
[10] http://www.nederlandserfgoeddigitaal.nl/
[11] http://www.europeana.eu/

Commitment met betrekking tot persistentie door langdurig bestaande instelling(en)
PIs zijn alleen persistent als een organisatie die persistentie garandeert. Daarvoor is op zijn minst nodig dat die organisatie zelf duurzaam is (grote erfgoedinstellingen zijn hiervoor bij uitstek geschikt, en hebben hier zelfs misschien wel een natuurlijke taak)

Geschikt voor toepassing door musea, bibliotheken en archieven
Mogelijk bestaan er aanvullende eisen aan algemene PI systemen vanuit het gebruik binnen de erfgoedsector.

Compatibel met Semantisch Web, in het bijzonder met Linked Open Data
Semantisch Web toepassingen zijn sterk vertegenwoordigd binnen de cases van de deelprojecten in CATCH Plus. Binnen de SW-gemeenschap worden objecten geïdentificeerd met behulp van z.g. http URIs (Uniform Resource Identifiers) . Verder bestaan er richtlijnen met betrekking tot het resolven van deze URIs binnen het Linked Open Data[12] initiatief. Het is van belang compatibel te blijven met LOD. Dit betekent onder meer dat persistente identifiers als resolvable http URIs moeten kunnen worden gerepresenteerd. Gekozen tekstrepresentaties voor identifiers moeten 'URL safe' zijn, omdat ze als (onderdeel van) URLs moeten kunnen worden gebruikt. (opmerking: URN-NBN voldoet hier niet aan, de DANS URN-NBN resolver wel). Ook het 'host'-gedeelte van de http URI moet een stabiele naam zijn, omdat die binnen de SW-gemeenschap als onderdeel van de identifier wordt gezien. Daarnaast is het een pré als een PI systeem het mogelijk maakt z.g. 'content negotiation' te implementeren: afhankelijk van het door de gebruiker gewenste type content kan de resolver een andere locatie URL teruggeven (voorbeeld: URL's voor html/skos/xml/json representaties van een concept uit een vocabulaire)

Organisatorisch

Betrouwbaar en redundant
Diensten rond persistente identifiers (in het bijzonder resolving) zijn essentieel voor het localiseren van data. Het is dus van het grootste belang, dat deze diensten betrouwbaar en snel zijn. Een PI resolver mag geen 'single point of failure' worden, dus redundantie is essentieel, bijvoorbeeld in de vorm van mirror sites.

Beperkte kosten van gebruik
Het is niet de bedoeling afhankelijk te worden van een specifieke leverancier, noch opgezadeld te zitten met periodieke licentiekosten. Alle gangbare PI technologiën zijn gebaseerd op open source software, dus van licentiekosten is geen sprake. Het is ook niet de bedoeling substantiële kosten te maken voor dienstverlening, of kosten per toegekende identifier. DOI kent kosten per identifier. Voor substantiële aantallen identifiers, zoals in CATCH Plus het geval is, valt DOI om deze reden als alternatief af.

Verdeling van verantwoordelijkheid en kosten van beheer
Het betreft hier een gezamenlijke onderneming waarbij bij voorzetting buiten CATCH Plus afspraken gemaakt gaan worden over verdeling van beheersinspanningen en bijbehorende kosten. De gekozen oplossing moet dit ondersteunen, bijvoorbeeld door expliciet vastleggen van gebruikers/administrators, gebruikersgroepen en hun rechten met betrekking tot bepaalde beheerstaken.

Identifier-beheer losgekoppeld van webserver-beheer en van hosting van identifiers
Het beheren van identifiers, het aanbieden van opslag en diensten rond identifiers en het beheren van webservers zijn heel verschillende taken, en moeten in principe dus kunnen worden uitgevoerd door verschillende personen, en mogelijk zelfs door verschillende organisaties. Identifiers moeten kunnen worden beheerd zonder tussenkomst van een systeem-/webserver-beheerder, degene die diensten rond identifiers verzorgt hoeft niet de eigenaar van die identifiers te zijn.

Consortiumbrede pool van persistente identifiers
Het ideaal is te komen tot één gedeelde pool van persistente identifiers voor het hele CATCH Plus consortium (en mogelijk breder). In principe mag die pool (mits redundant) gehost zijn op meerdere fysieke locaties.

Minimale beheersinspanning nodig
Het spreekt vanzelf dat beheerinstrumenten zo efficient en gebruikersvriendelijk mogelijk moeten zijn.

[12] http://linkeddata.org/

Afspraken over beleid mbt persistentie moeten zoveel mogelijk expliciet worden vastgelegd
Soms bieden PI systemen daarvoor zelfs ingebouwde mogelijkheden, zoals in het geval van ARK, waar een 'promise of stewardship' met iedere identifier kan worden geassocieerd.

Open voor andere erfgoedinstellingen
In principe is de binnen CATCH Plus ontwikkelde infrastructuur voor PI open voor gebruik door andere erfgoedinstellingen. Daarvoor moet duidelijk zijn onder welke voorwaarden dat kan, en wat er technisch en organisatorisch voor nodig is.

Institutionele onafhankelijkheid
Erfgoedinstelling kennen vele samenwerkingsverbanden. Invoering van persistente identifiers in het ene consortium mag samenwerking binnen andere verbanden niet in de weg staan.

Technisch

Granulariteit
Binnen CATCH Plus zijn identifiers nodig voor verschillende soorten objecten. Soms gaat het om grote aantallen, zoals bijvoorbeeld in het geval van geannoteerde tekstdocumenten, waarbij (semantische) annotaties aan segmenten binnen de tekst kunnen worden gekoppeld. Mogelijk moet ieder segment dan kunnen worden geïdentificeerd. Het kan hierbij gaan om honderden segmenten per document. Ofwel, dit stelt hoge eisen aan het snel kunnen resolven van grote aantallen identifiers (1 identifier per annotatie), ofwel dit vereist dat de resolver transparant om kan gaan met z.g. fragment identifiers (1 identifier per document).

Schaalbaarheid
Gezien het feit dat het toekomstig gebruik en omvang van de persistente identifier services nog niet bekend is, moet de gekozen oplossing schaalbaar zijn. Deze schaalbaarheid kent twee aspecten: schaalbaarheid met betrekking tot het aantal identifiers, en performance.
Middelen waarmee oplossingen schaalbaar worden gemaakt zijn:

- Hashing: methode die het mogelijk maakt identifiers over verschillende servers te verdelen en efficient te bepalen op welke server zich een bepaalde identifier bevindt.
- Caching: tijdelijke locale opslag van eerder 'geresolvede' identifiers.
- Replicatie: maakt het mogelijke meerdere sites in te richten voor dezelfde pool van identifiers.

Betrouwbaarheid
In technische zin kan dit onder meer worden bereikt door replicatie toe te passen (mirroring). Een variant is de resolver-oplossing zoals door DANS gerealiseerd: persistente identifiers worden door de aanbieders van collecties door middel van OAI-PMH[13] aangeboden en vervolgens periodiek geharvest door de centrale resolver.

Centraal geregistreerde Naming Authority
Het is om een aantal redenen belangrijk de Naming Authority voor een gegeven verzameling identifiers te registreren bij een centrale autoriteit. Dit maakt de resolver(s) voor een bepaalde identifier wereldwijd traceerbaar, het garandeert dat de idenfier (in combinatie met een uniek id voor de NA zelf) wereldwijd uniek is, en maakt het mogelijk redundantie en caching te regelen.

Metadata
Metadata hoort in principe thuis in een aparte catalogus en hoeft door het PI systeem dus maar minimaal te worden ondersteund. Zinvol zijn mogelijk: administrator/eigenaar van de identifier, contact-informatie, beschrijving van het geïdentificeerde object, of een statement met betrekking tot persistentie.

Eisen vanuit CATCH Plus applicatie-scenarios?
Een nadere analyse moet nog worden gemaakt. Hieruit moet duidelijkheid onstaan over te verwachten aantallen PIs, en soorten te identificeren objecten (user profiles, docs, concepten, annotaties, fysieke objecten, ..). Verder kunnen eisen met betrekking tot performance blijken te bestaan.

[13] http://www.openarchives.org/OAI/openarchivesprotocol.html

Persistente identifiers moeten groepsgewijs kunnen worden gewijzigd
Collecties en andere dataverzamelingen worden soms in hun geheel verplaatst. Dan moet het eenvoudig zijn alle bijbehorende persistente identifiers in één transactie te updaten. Dit geldt zowel voor verplaatsingen binnen als tussen erfgoedinstellingen.

Eisen aan client tools en software bibliotheken
Er kan op verschillende manieren van de diensten van het PI systeem gebruik worden gemaakt: via tools (applicaties), via web front-ends of applets, via client software bibliotheken, via web services, op protocol niveau (eigen protocol, zoals voor Handles, of http), vanuit standaard web browsers. Voor beheer door erfgoedmedewerkers is toegang door middel van applicaties en/of webtoegang tot diensten noodzakelijk (voor resolver beheer en gebruik). Voor software-ontwikkelaars binnen de diverse CATCH Plus deelprojecten en gezamenlijke diensten is toegang via software bibliotheken of web services van belang (voornamelijk voor resolver gebruik).
Tenslotte dient het mogelijk te zijn een persistente identifier te resolven met behulp van een standaard web browser. (dit impliceert dat er een html representatie beschikbaar moet zijn van de geassocieerde URL(s) of van de door die URL(s) aangewezen data)

Veiligheid en encryptie
In een enkel geval biedt een PI framework ondersteuning voor encryptie, digitale handtekeningen en server authenticatie (b.v. Handles). Vanuit de CATCH Plus gemeenschap moet duidelijk worden of hieraan behoefte is.

Oplossingsvoorstellen

Evaluatie van technische oplossingen

Mogelijke alternatieven voor diensten met betrekking tot identifiers zijn: uitsluitend gebruik maken van standaard webtechnologiën (http, URI, DNS), Handles, DOI, PURL, URN-NBN en ARK. Met uitzondering van URN-NBN zijn al deze alternatieve persistente identifiers te resolven door middel van een (http) URI. In al deze gevallen volgt die URI hetzelfde patroon van een ‘service request’: het bestaat uit een host-id, een service-id, en het persistente deel, de eigenlijke identifier.

Voorbeelden van dergelijke service requests:

http URI: http://www.beeldengeluid.nl/gtaa#Subject_aalscholvers
PURL: http://identifiers.erfgoed.nl/purl/vocabularies/iconclass/concept1821 of http://purl.org/vocabularies/iconclass/concept1821
Handle: http://identifiers.erfgoed.nl/hdl/1280.14/local_id_1821
ARK: http://identifiers.erfgoed.nl/ark:/128014/local_id_1821
URN-NBN: urn:nbn:nl-local_id_1821, plus resolver: www.persistent-identifier.nl

De daarbij behorende opdelingen in host-id, service-id en identifier:

Host id	*Service id*	*Geregistreerde Naming Authority*	*Eigenlijke identifier*
http://www.beeldengeluid.nl	-	www.beeldengeluid.nl	gtaa#Subject_aalscholvers
http://purl.org http://identifiers.erfgoed.nl	 purl	vocabularies -	iconclass/concept1821 iconclass/concept1821
http://identifiers.erfgoed.nl	hdl	1280.14	local_id_1821
http://identifiers.erfgoed.nl	ark:	128014	local_id_1821
-	-	nbn:nl-	local_id_1821

In principe is het dus mogelijk op één host met behulp van http meerdere typen identifiers te resolven.

DOI wordt veelvuldig gebruikt in de uitgeverswereld. Technisch gezien is het een specifieke toepassing van Handles, DOI’s zijn dan ook gewoon via Handle resolvers te resolven. Bij DOI moet worden betaald op basis van het aantal toegekende identifiers. Gezien het feit dat het binnen CATCH Plus gaat om mogelijk grote aantallen identifiers, laten we DOI op grond van kostenoverwegingen verder buiten beschouwing.

In principe is het mogelijk standaard URIs te gebruiken als persistente identifiers. Resolving wordt dan gerealiseerd door een combinatie van domeinnaam resolutie (via DNS) en ‘http redirection’. Het grote voordeel

is, dat op deze manier uitsluitend gebruik wordt gemaakt van beproefde, standaard webtechnologie. Echter, deze oplossing veronderstelt veel discipline bij het afspreken en toepassen van richtlijnen die persistentie moeten garanderen[14]. Bovendien bestaan noodzakelijke diensten voor het beheren van identifiers niet, hooguit kunnen beheerders van webservers enkele beheerstaken oplossen[15]. Verder is deze oplossing niet flexibel, omdat identifier-beheer, hosting van identifiers en webserver-beheer niet ontkoppeld kunnen worden. We concluderen dan ook, dat het gebruik van uitsluitend standaard webtechnologie niet toereikend is. Wel blijft de eis overeind, dat gekozen oplossingen voor persistente identifier diensten via standaard webtechnologie te gebruiken moeten zijn (via http en DNS, gerepresenteerd mbv http URIs, volgens cool URI en Linked Open Data best practices, URL-safe geëncodeerd).

Hoewel URN-NBN binnen Nederland als basis wordt gebruikt voor samenwerking aan infrastructuur voor persistente identifiers (o.m. door SURFShare, DANS en de KB) voldoet het niet aan een aantal van onze eisen: URN-NBN identifiers kennen geen URI representatie en zijn niet URL-safe. Naming Authoriteiten vallen per definitie samen met nationale bibliotheken, wat voor onze profielen niet flexibel genoeg is. Verder bestaat er geen bijgeleverde software om diensten rond URN-NBN's te realiseren. Deze diensten zijn een verantwoordelijkheid voor iedere nationale bibliotheek afzonderlijk. Voor de Nederlandse implementatie van URN-NBN diensten lijken betrouwbaarheid, performance en redundantie vooralsnog moeilijk te garanderen. Op dit moment laten we URN-NBN als oplossing dus maar buiten beschouwing.

Overblijvende technische oplossingen, die in min of meerdere mate voldoen aan onze wensen en eisen zijn PURL, Handle en ARK. PURL en Handle zijn beproefde platforms, die al meer dan een decennium op brede schaal worden toegepast. Handle heeft van deze twee met betrekking tot performance, schaalbaarheid, gedistribueerde opzet, authenticatie en authorisatie, security en beschikbare softwaretools en –bibliotheken veruit de beste papieren.

Ons voorstel is dan ook PURL en Handle identifiers beide te ondersteunen, waarbij Handle de basis-oplossing is. Voor wat betreft URN-NBN en ARK kan de CATCH Plus resolver eventueel 'redirecten' naar de betreffende externe resolvers.

Handles

Het Handle System (http://www.handle.net, van CNRI) is een wereldwijd veel toegepast gedistribueerd systeem voor identifiers en resolving. Sinds het midden van de jaren 90 wordt het gebruikt door een reeks universiteiten, nationale bibliotheken, overheidsinstellingen, rekencentra, en bedrijven. De grootste en bekendste gebruiker is de International DOI Foundation, die identifiers voor de internationale uitgeverswereld verzorgt.

Handles zijn identifiers met een heel eenvoudige syntax: "prefix/suffix", een voorbeeld is "10.1045/april2006-paskin". Het Handle System bestaat uit twee lagen: een globale dienst met de naam Global Handle Registry, en Local Handle Services. Iedere LHS komt overeen met een geregistreerde Naming Authority en heeft een unieke prefix (registratie van een prefix kost $50 per jaar). Het suffix gedeelte kan naar eigen inzicht van die NA worden gebruikt. De Global Handle Registry bevat speciale Handles, die prefixes met de locaties van Local Handle Services associëren. Local Handle Services beheren en resolven hun eigen identifiers. Via de Global Handle Registry heeft iedere LHS in maximaal twee stappen toegang tot alle Handles wereldwijd.

Het Handle Systeem is speciaal ontworpen met het oog op performance en schaalbaarheid. Het kent ingebouwde mogelijkheden voor Local Handle Services om mirrors in te richten, om identifiers over meerdere servers te verdelen en om identifier informatie tijdelijk locaal op te slaan (caching).

Handles kunnen worden geresolved via een eigen handle protocol, daarnaast zijn alle Handles te resolven via een http proxy server (b.v. via http://hdl.handle.net/4263537/5555). De Handle software ondersteunt ook resolven via http door locale Handle services (b.v. http://identifiers.erfgoed.nl/hdl/4263537/5555).

Tenslotte, binnenkort gaan Handles ook ondersteuning bieden voor z.g. fragment identifiers (deel van een URL achter een #-teken, dat wordt gebruikt om een gedeelte van een resource aan te geven). Het Handle System stript dan eerst de fragment identifier van de URL, resolved het restant, en plakt de fragment id weer achter de resultaten. Dit is een belangrijke feature voor het gebruik van identifiers voor annotaties, die aan delen van digitale objecten zijn gekoppeld.

---

[14] http://www.w3.org/TR/cooluris/
[15] http://www.w3.org/TR/swbp-vocab-pub/

Profielen

In de (zeer lezenswaardige) documenten van het Australische PILIN project[16] worden vier ‘management profielen’ met betrekking tot het beheer van persistente identifiers geïntroduceerd. Onderscheid tussen deze profielen wordt gemaakt langs de dimensies ‘hosting van identifiers’ (centraal versus eigen hosting) en ‘beheer van identifiers’ (eigen beheer versus gezamenlijk beheer). Een citaat:

“The default profile is *devolved management*, where each party manages their own identifiers and identifier systems. (own hosting, exclusive management)

Centralised systems introduce an economy of scale, and relieve parties of administrative burdens; but they also take away much of the ownership of the identifiers being pooled into the centralised system. (central hosting, shared management)

Autonomous systems address this problem by uncoupling hosting from identifier management. (central hosting, exclusive management)

Federation addresses the problem in a different way, restoring ownership through a shared consortium—but this introduces its own administrative burden at the consortium level. (own –federated- hosting, shared management)”

Het hangt sterk van de voorkeuren en randvoorwaarden van de deelnemende erfgoedinstellingen af welk profiel voor CATCH Plus (en mogelijkerwijs NED! en Europeana) het meest geschikt is. Vooralsnog gaan we ervan uit dat de deelnemers in CATCH Plus de ambitie hebben als consortium samen stappen te nemen, dus laten we ‘devolved management’ buiten beschouwing.

Hieronder worden drie alternatieve voorstellen gepresenteerd, die de profielen ‘autonomous systems’, ‘federation’ en ‘centralised systems’ volgen.

Alternatief 1: autonomous systems

Bij dit model vinden de opslag van identifiers en het aanbieden van diensten centraal in CATCH Plus (of voor het Nederlands erfgoed) plaats. Het beheer van de identifiers wordt echter in principe gedaan door ieder van de deelnemende erfgoedinstellingen zelf. Het voorstel is te beginnen met één museum (Rijksmuseum Amsterdam), één bibliotheek (Koninklijke Bibliotheek) en één archief (Beeld en Geluid). Anderen kunnen later aansluiten met hun eigen pool van identifiers of bij één van de drie starters.

Concreet betekent dit, dat er drie “naming authorities” komen, ieder met een eigen geregistreerde Handle prefix. Het CATCH Plus projectbureau regelt centrale hosting en identifier services voor de drie Local Handle Systems. Er dient daarbij gezorgd te worden voor redundantie, betrouwbaarheid, performance en persistentie van de diensten. Daartoe wordt gezocht naar geschikte partners. In principe zijn er zo drie resolvers (we noemen ze even http://identifiers.musea.erfgoed.nl, http://identifiers.bibliotheken.erfgoed.nl, en http://identifiers.archieven.erfgoed.nl). Ieder van die drie kan Handles resolven (b.v. http://identifiers.musea.erfgoed.nl/hdl/4263537/5555) en desgewenst PURL identifiers.

Ieder van de naming authorities stelt een ‘identifier manager’ aan, die verantwoordelijk is voor het correct beheren van de eigen persistente identifiers. In principe kan iedere instelling zijn eigen beleid voeren met betrekking tot naamgeving en beheer.

Alternatief 2: federation

Bij dit model treden de deelnemers meer als consortium op: er wordt een gezamenlijke pool van identifiers in het leven geroepen volgens een gedeeld beleid met betrekking tot beheer en naamgeving. Deze totale pool van identifiers wordt door meerdere partners gehost: iedere partner host naast zijn eigen identifiers ook die van de anderen, waardoor redundantie ontstaat.

In dit geval komt het consortium overeen met één naming authority, met één geregistreerde Handle prefix. Er wordt slechts één Local Handle System in het leven geroepen, die bijvoorbeeld bestaat uit een drietal Local Handle sites (Koninklijke Bibliotheek, Beeld en Geluid, Rijksmuseum Amsterdam), die mirrors van elkaar vormen. Op deze manier ontstaat één gedeelde resolver, bijvoorbeeld met de naam http://identifiers.erfgoed.nl. Wederom kunnen Handles en PURL identifiers worden geresolved.

[16] https://www.pilin.net.au/Project_Documents/Community_Guidelines/Guidelines.htm

Evenals bij het bovengenoemde alternatief worden ook hier drie “identifier managers” aangesteld, echter nu is er sprake van een onderling afgesproken beleid. Overigens is als onderdeel van de gedeelde policies nog steeds mogelijk in de Handle suffix onderscheid te maken naar erfgoedinstelling, hoewel dit misschien minder gewenst is. (http://identifiers.erfgoed.nl/hdl/12345/rma-678)

Alternatief 3: centralised systems

Hierbij worden hosting en dienstverlening en een deel van de beheer-policies uitbesteed aan een centrale autoriteit, waarbij de belangen van de deelnemende partijen worden behartigd door vertegenwoordiging in een overkoepelend bestuursorgaan. In concreto, momenteel wordt ten behoeve van eScience gewerkt aan de opbouw van een consortium van grote academische rekencentra (momenteel met beoogde partners in Duitsland, Finland en Nederland). Dit consortium gaat voor Europa Handle services aanbieden op een autonome en redundante manier, en bovendien een mirror draaien voor de Global Handle Repository. Hierover zijn reeds principe-afspraken gemaakt met Amerikaanse CNRI.
Deze oplossing is vergelijkbaar met DOI, waar de uitgeverswereld toegevoegde diensten aanbiedt bovenop het Handle Systeem, met dien verstande, dat hier een ander businessmodel gaat worden gehanteerd, waarbij niet per toegekende identifier moet worden afgerekend.

Erfgoed-identifiers zouden in dit geval worden geresolved door iets als http://escience.eu/hdl of een Nederlandse equivalent van http://handle.gwdg.de . Zo’n identifier zou er dan bijvoorbeeld uit kunnen zien als: http://escience.eu/hdl/[vaste-prefix]/CH-NL-[identifier] .

Over registratie van identifier managers van de erfgoedinstellingen zouden in dit geval afspraken moeten worden gemaakt met de centrale autoriteit.

Aanvullende keuzen

Naamgeving
Met betrekking tot het kiezen van tekststrings voor identifiers wordt aanbevolen geen elementen op te nemen die in de loop van de tijd mogelijk gaan veranderen. Dus geen semantiek en geen technologie-afhankelijkheden. Verder dient de eigenlijke locale identifier voorafgegaan te worden door een unieke id van een geregistreerde Naming Authority (Handle prefix, ARK Name Assigning Authority Number, PURL padnaam, URN-NBN ISO country code).

PURL resolving
Hiervoor zijn er twee manieren van ondersteuning mogelijk. Ten eerste is het mogelijk een eigen domain op purl.org te regelen, en daaronder, centraal, PI-URL associaties te beheren. De tweede optie is een eigen locale PURL resolver in te richten voor onze PI-URL associaties, en die globaal te ‘registreren’ bij purl.org (dmv een ‘partial redirect’). In dit laatste geval kan http://purl.org/erfgoed-nl/XXX eerst worden geresolved naar http://identifiers.erfgoed.nl/purl/XXX , en daar vervolgens naar de daarmee geassocieerde URLs. Deze laatste optie kan worden ingevoerd als dat qua omvang en performance wenselijk wordt.

URN-NBN ondersteuning
Het zou mogelijk moeten zijn alles wat begint met http://identifiers.erfgoed.nl/urn:nbn:nl- te redirecten naar bijvoorbeeld de DANS resolver http://www.persistent-identifier.nl . Wel moet gezorgd worden dat alles “URL safe” gebeurt.

Versie-beheer
Een algemene identifier voor het object verwijst naar de laatste versie, eventueel kunnen aparte identifiers worden gebruikt voor iedere versie die apart te localiseren moet zijn.

Metadata
Het streven is slechts minimale metadata te associëren met identifiers, namelijk alleen die metadata die ten dienste staat aan identificeren en localiseren van objecten. Beschrijvende metadata hoort thuis in catalogi of andere metadata repositories.

Te realiseren diensten
We streven ernaar voor identifiers.erfgoed.nl de volgende extra diensten te realiseren

- Automatische generatie en/of validatie van nieuwe identifiers
- Content negotiation: we streven ernaar service requests voor Handle (en eventueel PURL) content negotiation te laten ondersteunen. Voor een gegeven identifier moeten dan meerdere geassocieerde URLs kunnen bestaan, die te onderscheiden zijn naar type. Handles hebben daartoe de technische voorzieningen, voor PURLs is dit waarschijnlijk alleen te realiseren door een aangepaste locale PURL resolver te gebruiken.
- Groepsgewijs updaten van identifiers

Actiepunten/plan van aanpak

Voor het succesvol invoeren van persistente identifiers in het CATCH Plus consortium is bewustwording van de problematiek en inzicht in de mogelijke oplossingen nodig. Verder moeten door de samenwerkende erfgoedinstellingen in onderling overleg beleidskeuzen worden gemaakt. Daarvoor is meer tijd nodig dan in de eerste fase van CATCH Plus beschikbaar is. Daarom stellen we voor een 'beleids-traject' en een 'implementatie-traject' parallel uit te voeren, en bij de implementatie zoveel mogelijk beleidskeuzen open te houden.

Beleids-traject

- feedback van partner-projecten en –organisaties in dit voorstel verwerken (van NED!, Europeana, VU)
- Mini workshop (reeks?) met eerstbetrokken CATCH Plus erfgoedinstellingen (B en G, KB, Naturalis, RMA) en andere CATCH Plus geïnteresseerden. Doelen voor deze workshop(s):
    - Maken beleidskeuzen, naar aanleiding van de lijst in dit document
    - Kiezen van een profiel met betrekking tot hosting en beheer.
    - Support voor technische keuzen verkrijgen.
- Periodieke evaluatie van het implementatie-traject.

Implementatie-traject voor 2009

- Schrijf technische specs
- Vind daarbij een implementator
- Regel initiële hosting van de persistente identifiers en bijbehorende diensten
- Implementeer/configureer standaard Handle services, inclusief eventuele registratie van Naming Authority(ies)
- Implementeer extra PI diensten
- (optioneel) Implementeer locale PURL resolving en bijbehorende centrale registratie.
- Zet een eerste versie van een gebruikersadministratie op
- Ken persistente identifiers voor vocabulaire-elementen toe
- Voer persistente identifiers in voor 1 of 2 pilot collecties. Benodigde stappen:
    - 1. introduceer tabel die proprietaire id koppelt aan een persistente id. Leg voor die proprietaire id ook informatiesysteem, locaal veld en datatype van dat veld vast. 2. voeg aan proprietair systeem een extra veld toe met een persistente identifier, naast de proprietaire identifier.
    - Zorg dat geïdentificeerde objecten zinvol via een URL zijn te retrieven. (voorbeeld: pi → url → immix record id → database record → DC/XML/RDF representatie via http
    - Regel dat bestaande verwijzingen (mede) via persistente identifiers verlopen
    - Start met direct toekennen van PI's aan nieuwe objecten door collectiebeheerders