The ARK Identifier Scheme: Lessons Learnt at the BnF and Questions Yet Unanswered

Sébastien Peyrard, Jean-Philippe Tramoni, John Kunze

The ARK Identifier Scheme: Lessons Learnt at the BnF and Questions Yet Unanswered

Sébastien Peyrard
BnF, France
sebastien.peyrard@bnf.fr

Jean-Philippe Tramoni
BnF, France
jean-philippe.tramoni@bnf.fr

John Kunze
California Digital Library, USA
jak@ucop.edu

Abstract

The Bibliothèque nationale de France (BnF) looks back at lessons learnt over eight years of implementing persistent identifiers (ARKs). While persistent identification is still a relatively young field, this is enough time to gain practical experience, and to conduct a meaningful gap analysis between what is and what should be, especially in a semantic web context. That analysis has exposed important issues concerning best practices and compliance with existing standards.

Keywords: Archival Resource Key; persistent identifiers; web of data; linked data.

Introduction

"Eternity is a very long time, especially towards the end." W. Allen[1]

When considering persistent identifiers, one tends to focus on two ends of the timeline: the immediate near term (at the initial implementation stage) and the very long term, the latter often being too abstract to act on directly. After eight years of implementation experience and almost 20 million ARKs assigned, the BnF now takes the opportunity to look back. This article explores what issues have to be considered during the lifespan of persistent identifiers, in this case ARKs. It also touches on the ARK standard: this 13-year-old standard might benefit from clarification or modification. At a time when institutions are diving into linked data and appear as key stakeholders in the web of data, we believe persistent identifiers have a key role in supporting trustworthy and stable bridges across data silos.

1. The ARK identifier scheme: overview

ARK identifiers have been introduced in various articles and web resources (CDL, 2013) (Kunze, 2003). This section summarizes only enough to make the rest easily understandable.

1.1. Purpose and aim

The ARK standard addresses the same issues as other persistent identification schemes. Although anyone can use them, and there are about 270 organizations currently registered (CDL, 2014), ARKs have been most popular with heritage institutions. These institutions are usually tasked with indefinite retention of content, well beyond expected lifetimes of commercial institutions, and where the perspective is set on the very long term.

ARKs have a very conservative approach to persistent identification. Like URNs and DOIs, ARKs are designed to be independent of DNS and the HTTP protocol; however, they are also designed to work directly in today's web environment URLs, by specifying that the hosting arrangement does not affect identity. For example, these ARKs identify the same resource:

- http://gallica.bnf.fr/ark:/12148/bpt6k5834013m
- http://bnf.example.org/ark:/12148/bpt6k5834013m

[1] http://edition.cnn.com/2006/WORLD/asiapcf/07/04/talkasia.hawking.script

- ark:/12148/bpt6k5834013m

The last of these (with no hostname) is the *core immutable identifier*.

1.2. Anatomy

The base ARK name is typically a completely opaque (meaningless) identifier in order to drastically reduce any pressure to change the identifier string over the long term. For example,

- http://gallica.bnf.fr/ark:/12148/bpt6k5834013m
- This sort of base name is often extended with a qualifier that may be less opaque, as in
- http://gallica.bnf.fr/ark:/12148/bpt6k5834013m/f19.highres

An actionable ARK (an ARK that works in today's web) has three main parts.

- The *core immutable identifier* itself is mandatory and is designed to be globally unambiguous, persistent and opaque. To that end, it has a structure proceeding from the most general to the most specific (left to right):
  - the *identifier scheme* ("ark:/"), a label that is easy to find by simple text miners;
  - the *Name Assigning Authority* (NAA), which has a 5-to-9 digit NAA Number (NAAN) for opacity. NAAN uniqueness is guaranteed via a registry[2] based at the California Digital Library (CDL);
  - the *ARK name* itself, which should be opaque and is assigned by the NAA; if independent ARK name assignments are performed within a single NAA, the NAA often designates sub-naming authorities corresponding to short prefixes for the ARK name, to ensure ARK names uniqueness.
- The *Name Mapping Authority* (NMA), which enables the identifier to resolve to a resource. The NMA is implemented with a Name Mapping Authority Hostport (NMAH), which in today's web environment is usually an HTTP server. This part can change over the long term, which is why it is optional. Here for example the NMAH is "http://gallica.bnf.fr".
- The optional *qualifier part*, which enables extra services provided by the NMAH using the standard ARK reserved characters "." and "/". At BnF they are often used as follows.
  - Naming sub-parts of a resource (e.g. a specific page in a digitized book). This is achieved by *hierarchy qualifiers* beginning with "/" (/f19 in the example).
  - Naming variants or services of the resource (e.g. a specific version in the lifecycle of a digitized book, or the thumbnail of a given image). This is achieved by *variant qualifiers* beginning with "." (.highres in the example)

1.3. Using ARKs

ARKs raise many of the same issues as other persistent identification schemes.

- **Institutional commitment and policy**. Persistent identification is not a technical problem. It will only work if an institution commits to ensure persistence and global uniqueness over the long term. There needs to be a clearly articulated stewardship policy.
- **Assignment procedures**. Clearly articulated procedures are also required to ensure that assignments are unique and consistently applied to defined resource types. Decisions to be made comprise what ARKs are identifying, which resources are considered to deserve separate ARKs, and which resources should be considered variants of the same ARK.
- **Resolution**. One or more NMAHs are needed to resolve ARKs, each NMA defining a level of service provided with the ARKs. Reliable resolution allows reliable *citation*.

[2] The NAAN registry can be accessed at http://www.cdlib.org/uc3/naan_registry.txt.

ARKs also offer two ways of supporting linked data. Besides using content negotiation, ARK end-users may instead append suffixes, called *inflections*, to gain access to services related to a resource, but without requiring them to remember whole new identifiers. For example,

- http://texashistory.unt.edu/ark:/67531/metapth346793/ *(ARK for the resource)*
- http://texashistory.unt.edu/ark:/67531/metapth346793/? *(its metadata)*
- http://texashistory.unt.edu/ark:/67531/metapth346793/?? *(the NMA's commitment)*

By itself an ARK should lead to the resource (object). Appending a single "?" should lead to the resource's metadata (Kunze, 2010) and appending "??" should lead to metadata describing the kind of persistence to expect. In the current archival environment, the latter is critical for indicating when a resource is truly invariant, or subject to correction, or is a growing resource. As an alternative to content negotiation, ARK inflections are easier to use and more precise. Inflections are not as easy to support, however, with the Tomcat-based web services at BnF.

2. A brief history of ARKs at the BnF

2.1. Adoption and initial implementation of the ARK identifier scheme

In 2006, the BnF conducted a risk-driven requirements analysis to adopt the ARK persistent identification scheme. Two core requirements used for selection criteria were (1) *financial independence* of the NAA: identifiers subject to a fee, such as DOIs, were discarded and (2) *technical independence* of the naming authority (since identifiers had to be directly integrated into our in-house Information Systems): identifiers relying on installing special-purpose software, such as Handles, or on external services, such as PURLs, were discarded. BnF needed *stable, location-independent URLs*, which do not redirect to temporary URLs (avoiding the overhead of managing an endlessly increasing number of redirects).

URNs also fit our criteria fairly well, but the ARK specification addressed some areas more precisely than URNs, such as the definition of a persistence policy, and additional services on a particular resource in a web context (through the use of qualifiers). Like the URN scheme, the ARK scheme does not mandate use of one particular vendor or service for its identifiers. Unlike URNs, DOIs, and Handles, however, ARKs also do not mandate use of one well-known DNS resolution starting point, so ARKs can be implemented directly on a local web server. While some consider this a weakness, citing the "inherent" fragility of DNS names, their argument usually suggests using dx.doi.org, handle.net, or n2t.net instead; the logical flaw is that these are DNS names too, and we note that none of them are as long-lived as bnf.fr. The bottom line is that ARKs are implementable with the simplest of technologies, and they do not require a special-purpose global infrastructure uniquely built for their own scheme.

At this stage, ARKs were defined for two distinct types of resources: **digitized documents**, available in the digital library Gallica – using http://gallica.bnf.fr as NMAH and **catalogue records**, which needed identification for exchange with BnF's OAI repositories – using http://catalogue.bnf.fr as NMAH.

For both NMAHs, we defined an initial complete set of qualifiers to name subparts and variants. As an illustration, in gallica.bnf.fr, we defined qualifiers to name the pages of a book (e.g. http://gallica.bnf.fr/ark:/12148/bpt6k5834013m/f10 to name page number 10 in the digitized document, /f10n5 to name the set of pages 10 to 14), and qualifiers to invoke variants of a book or a page (http://gallica.bnf.fr/ark:/12148/bpt6k5834013m/f10.highres, .medres, .lowres and .thumbnail for the different resolutions of the same page; .text to access the OCR for a particular page, .vocal to access the sound version of the same page). For the main catalogue, qualifiers were used to name distinct formats of the same record.

More details about the initial approach and the first implementation choices are available in (Bermès, 2006). During the eight following years, ARKs became the lingua franca across the institution, and their use expanded to new areas.

2.2. Fostering ARK identifiers: new resources, new clients

Since 2006 BnF has expanded its initial use of ARKs for two different purposes:

- Identifying descriptive records in order to manage them in our **OAI** repositories, and more recently, in our data.bnf.fr **linked data** services. This led to assigning ARKs to EAD finding aids, manuscript illumination records, museographic descriptions.
- Preserving digital documents. In 2010 our preservation repository, SPAR (Scalable Preservation and Archiving Repository), went operational. As each Information Package had to have a persistent identifier, SPAR played the role of an ARK assigner whenever there was no pre-existing ARK assigned to the ingested document.

These different resource sets had different scales and creation workflows, which made it very difficult to have a single ARK assignment procedure. The most central assigner is SPAR, but it is only for digital documents (not descriptive records) and it was rolled out after the assignment channels for mass digitization were operational and optimized, which led to path dependence. On the descriptive records end, some databases had much smaller datasets than the 15 million records of the catalogue, which made semi-automated assignment procedures more suitable.

In the end, ARKs were assigned using three different means.

- **Automated, based on an existing number**: used for our two legacy systems (Gallica and our catalogue records), and for our finding aids database. Our large datasets have pre-existing reliable numeric ids that we can "dress" as ARKs. E.g. the record n°32915216 from the main catalogue had the "c" sub-naming authority for descriptive resources, and the "b" 2nd level sub-naming authority for records from the main catalogue. Thus, 32915216 became ark:/12148/cb32915216j (with the addition of a final check character).
- **Automated, independent of any number**: used for medium to large datasets with no reusable id (because significant or incompatible with the ARK structure). Our preservation repository, SPAR, automatically assigns an ARK upon ingest. E.g. ark:/12148/bc6p01zndd assigned to a web archiving container file, indicates (to BnF staff) that assignment was routed to sub-naming authority "b" (digital content) and to repository "c6p0", a 2nd–level sub-naming authority that takes care of uniqueness at repository level.
- **Semi-automated**: with a list of ARKs that curators assign to resources (one spreadsheet per sub-naming authority), this is used for very small datasets. It meant defining a sub-naming authority per database to guarantee uniqueness. E.g. ark:/12148/cdt9x5ww identifies a book binding description. As a descriptive record, assignment was routed to sub-naming authority "c", then to 2nd-level sub-naming authority "dt9x" for book binding.

On the access side, as new services were being built upon new resources, several ARK NMAHs could be used simultaneously for the same resource[3]. For instance, the same catalogue record can be displayed in the main catalogue, which delivers a "full" but isolated record in traditional formats, and also in data.bnf.fr, which provides the RDF view of this record, but displays it in an enriched landing page that aggregates related resources. The difference is obvious for authority records, which can be seen, for instance, between these two ARKs:

http://catalogue.bnf.fr/ark:/12148/cb118905823

http://data.bnf.fr/ark:/12148/cb118905823[4]

[3] This might be considered a risky practice, as with several NMAHs for the same ARK identifier, you need to know all the NMAHs of a particular resource to have a complete view of it. We addressed this problem by defining a default NMAH for a given resource that is considered the "master" view for such a resource. For instance, http://catalogue.bnf.fr (main catalogue) is the default for bibliographic descriptions. A strength of potentially distinct NMAHs for a single ARK is that it forces one to dissociate the resource from the current application providing access to it, which forces one to adopt a long term perspective.

[4] As of 2014, July, data.bnf.fr accounts for only 60% of the catalogue data. Therefore, 40% of the ARKs in the main catalogue are not (yet) in data.bnf.fr.

2.3. At international scale: backing up the ARK registry

The NAAN registry maintained by the CDL described in §1.2 is a cornerstone for the viability of ARKs, because the centralized registration of NAAs ensures the uniqueness of each NAA Number (NAAN). To this end, it was important to guarantee its persistence over the long term, which led to registry mirroring arrangements with the US National Library of Medicine (NLM) and the BnF. From the BnF point of view, it meant formalizing a partnership with the CDL with a Memorandum of Understanding. As this MoU had to be signed off by the president of the BnF, it had the beneficial side-effect of securing institutional commitment for ARK identifiers from top-level management.

3. Implementation gap analysis: Consolidating ARK curation at the BnF

The previous section describes how ARKs gained momentum at the BnF and were progressively applied for different purposes and resource types beyond the originally envisioned use cases. This led to a wide variety of implementation choices and management rules, and consequently a call for centralized policy and harmonization. A gap analysis was conducted in 2014 to address this question in a systematic fashion. It consisted of summarizing the lessons learnt and problems encountered over the past 8 years, and then organizing those lessons around the following focal areas: functional, organizational or technical issues, qualifier implementation questions, policy descriptions, and compliance with standards. Those focal areas are described in separate subparts of §3 and §4.

The next subsection summarizes the issues uncovered by the gap analysis. Most of them are not complicated technical issues, but rather simple observations that we think would likely be made by any organization similar to BnF after 8 years of managing persistent identifiers.

3.1. Organizational issues

A persistent identifier and its policy should outlive its initial implementers. Obvious as this statement sounds, its direct implications are not readily apparent in the early implementation stages. It requires continuous improvement and refinement of the identifier policy and uses, which must remain stable while accommodating new and evolving uses and needs. This prevents identifiers from falling into obsolescence or disgrace, with a decrease in perceived relevance or visibility. Neither must they become "over-used"; frequent or casual assignment leads to misuse. A disciplined approach to organization and communication are key factors to sustainability.

In eight years, there has been a good deal of staff turnover in the ARK BnF expert team. Only one person from the original seven-member team remains. What's more, as ARK use expanded to new areas (as addressed in §2), its audience got much wider than the original team. This includes **library curators** that use, or might use, ARKs to cite resources; **digital object curators**, that handle the lifecycle of the object, including identification and access; **web application managers**, on the IT and librarian sides; **linked data experts**, especially for the data.bnf.fr project. As a result the communication and documentation had to be adapted for the larger audience, which needed to be aware of policy and key curation issues without necessarily understanding all the details.

Our "ARK consolidation approach" had two organizational phases.

Communicate: gather all the users, train them in the main underlying concepts of persistent identifiers, common misconceptions about them and best practices, and mandate two "reference ARK coordinators" – one on the IT and one on the librarian side.

Set up targeted working groups, led by the "ARK coordinators", these focused on specific resource types or applications, reducing the identified gaps and addressing new needs.

3.2. Functional gap analysis

The functional gap analysis itself revealed many areas for improvement in our persistent identifier services, particularly for resolution and associated services[5].

- Some applications do not create resolvable ARKs, but only record them as metadata.
- Whenever a resource is not available in the ARK-aware URL, there is only a 404 or 403 browser response, which should be replaced by one of the following more explicit statements: 1) *Resource not found* – this is an incorrect URL and no resource has ever been available at this URL; 2) *Resource deleted* – the resource was there, but it was deleted; in this case, provide core metadata and if possible the reason for the deletion; 3) *Access disallowed* in this context; as with deletion, one should provide core metadata and if possible the reason of the withdrawal (e.g. copyright status).
- Across some applications there are obsolete or inconsistent ARK redirects. E.g. an old test version of the digital library, gallica2.bnf.fr, no longer redirects to gallica.bnf.fr.

In all these cases, our minimum baseline service is clearly not achieved. Our first goal is therefore simple but attainable: define BnF "ARK core services" that any persistent-id aware application should comply with, namely,

- Provide access to the object behind the ARK
- In case of object unavailability, provide metadata to understand what was there and why access is no longer possible.
- Set up a generic process for updating redirects at the level of the BnF "ARK coordinators".

3.3. Refining the identification and persistence policies

When ARKs were first implemented, we had an unclear view of what stewardship promise we could return with identifiers. Therefore we ended up with a very high-level statement[6]:

- No identifier re-assignment;
- Identifier string policy: opaque strings, no vowels, use of a final check character;
- Persistence policy: guaranteed, but needs to be refined in the future; the form of the underlying resource can change to ensure its persistence (e.g. format migration).

With almost a decade of experience managing ARK identifiers, digital preservation objects (PREMIS Maintenance Activity, 2012), and alignments between our catalogue records and other linked data sources, we can see possibilities for differentiated persistence policies.

- For a digital document that we preserve, our aim is to keep the information content stable and accessible and useable to end-users. This means permanent access with stable content.
- For a catalogue record, the information content can be updated as the catalogue record is corrected, enriched, updated, etc. This means permanent access with somewhat more dynamic content.
- For an archival records document, the identifier will be maintained but the content may be suppressed for legal reasons. In this case, we provide a "tombstone" with the metadata and reasons for the object unavailability.

The BnF is currently considering formalizing these policies in a systematic way.

[5] This analysis is limited to ARK implementation. Time permitting, BnF could have studied additional identifier systems to get ideas for improvement, however the ARK scheme, being built upon experience with other schemes, was already a leading choice, so a broader study was not considered a priority.
[6] http://gallica.bnf.fr/ark:/12148/btv1b8451622d.policy

3.4. Refining the qualifier implementation

One issue we have to deal with is proliferation of identifier qualifiers (introduced in §1.2), in response to which we decided to create a consistent qualifier policy. From the most generic service to the most specific, we see three tiers of qualifiers.

- *Generic qualifiers.* Applicable to any resource, these are qualifiers providing a description of the resource (.description), its persistence policy (.policy), and potentially a qualifier revealing the sub-parts and variants available for the object.
- *Content-type-dependent qualifiers.* For digitized documents, you can use generic display resolution variants (thumbnails, low, medium or high resolution). For descriptive records, you can use generic metadata formats (RDF, XML…). The list of possible qualifiers can be maintained independently of any application.
- *Application-specific qualifiers.* These are specific to a particular NMAH.

We also consolidated our policy about when it is appropriate to define a new qualifier, due to two considerations revealed in the gap analysis. The first has to do with *querying vs. citations*. Variant qualifiers are *not* a query language, but do allow citation of services that one considers "persistent" and relevant from an end-user point of view. In that light, http://gallica.bnf.fr/ark:/12148/bpt6k65581775.r=food, in particular, the ".r=" qualifier raises a red flag. This qualifier can be viewed as a way to search for a word in a digitized document; but ARK qualifiers are intended to refer to the document, not to "look" into documents. It can also be viewed as a way to act upon a document by returning it (from BnF) "with highlights" added (here on the word "food"). This use case could comply with ARK qualifiers, but the side-effects could be distracting if not misleading. Unfortunately, it is easy to do accidentally; if a user previously searched for a word in a document before copying and pasting the URL, it will include the "r=word" qualifier. In the end, this creates a reference to a document with highlights, whilst most of the time all the user wants to do is refer to the document without them. This means that, in most cases, revealing such parameters is *not* recommended for persistent URLs.

A second consideration is *technical vs. non-technical* qualifiers. Any qualifier that concerns a detail of implementation, technology, or a temporary information object should not be expressed in the URI. Unlike the "ARK name" part, qualifiers are not meant to be long-term persistent. However, their stability and maintenance is important for the perceived trustworthiness of the service, and it is costly. Supporting the aforementioned .r= qualifier has a cost, as the syntax for searching for several words ".r=word1+word2+wordn" has to be maintained over re-implementations.

As a result of this gap analysis, the BnF intends to raise awareness of good practices among ARK users (developers and web application managers) and to formalize a general best practices document. A list of qualifiers will be created and maintained for the three aforementioned levels.

3.5. Technical issues: consolidating the technical framework

From its first implementation, the ARK resolver at the BnF had to meet two basic requirements: complying with the security policy of the IT operations service and managing the increasing flow of network requests.

Initially, the ARK resolver was a part of a general-purpose document viewer application. For each domain-specific application, every incoming URI including an /ark:/ pattern had to be detected by an HTTP reverse proxy and redirected toward this viewer application. The ARK resolver had to analyze the ARK identifier and the request, change it to a domain-specific format, and then forward the request for processing to the domain-specific application. These applications were hosted on multiple servers using virtual IP and load-balancing in order to share the load between these servers. This architecture had some shortcomings. First, the use of a reverse proxy conflicted with the IT operations requirements. Second, to detect, change and

redirect the requests, the ARK resolver had to implement some domain-specific rules. This was dangerous for the security and maintainability of the whole system.

After this first architecture was in operation for two years, it was agreed to define a new system that would be more generic, parameterized, and scalable. The multi-server load-balancing system was kept, but three modules were added.

a) A domain-specific module that checks if the incoming request is in the scope of the domain, and if not, sends it to the *ARK redirection module*. This filter module is generic but uses domain-specific patterns to verify incoming requests before they go to module b).

b) Domain-specific sub-modules analyze the request, and if necessary, reformat it according to the domain's requirements before transferring it to the domain-specific application.

c) The *ARK redirection module* is able to analyze the ARK identifier and the incoming request and then forward the request for processing to the domain-specific application. The redirection rules are parameters defined in an XML file.

The new document viewer application is now leaner because it does not handle the resolution of ARK requests. This task has been distributed between the generic redirection filter, the specific reformatting filters, and the centralized ARK redirection module. The workload of this redirection module is lower since many of the incoming requests are going directly to a domain-specific application that can resolve the ARK identifier.

Three years later, new requirements came out in parallel with new developments of the Gallica viewer module. Some tools were implemented to manage ARK identifiers and qualifiers, which are now defined by a configuration file. The processing of ARK qualifiers gained leverage by becoming more generic, which made them easier to use in the Gallica API. The ARK redirection module was enhanced by migrating the old redirection rules to mapping tables stored in a database. That module is also using a copy of the ARK NAAN registry that is mirrored at regular intervals from the NAAN registry at the CDL. The new architecture is summarized in Figure 1.

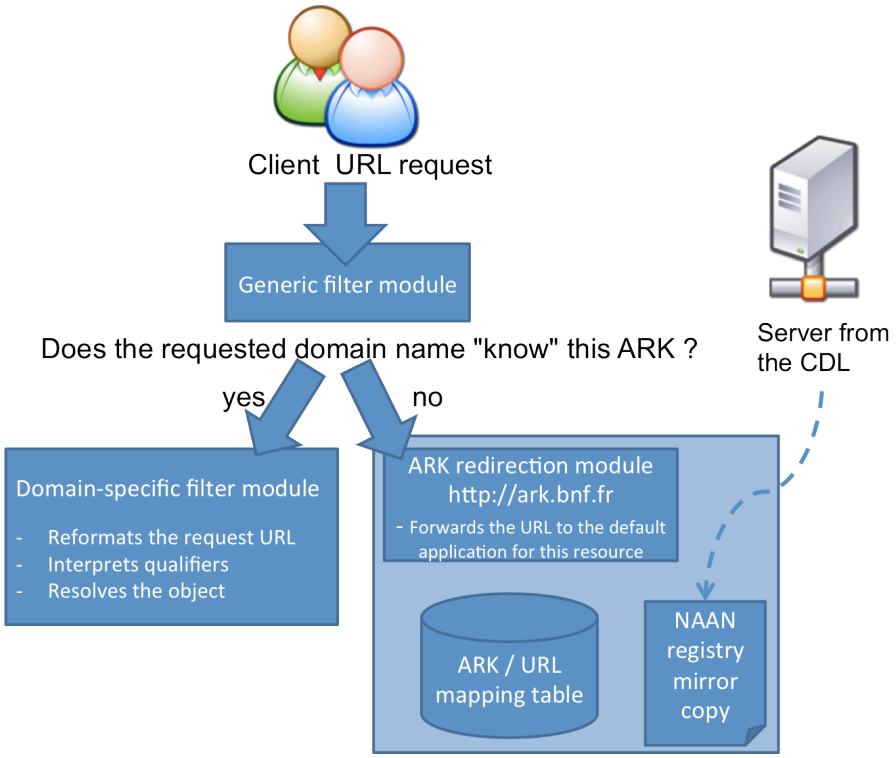

FIG. 1. BnF ARK resolver architecture

The ARK minting process, functional aspects of which are outlined in §2.2, has followed a similar evolution. Initially, ARK creation was completely delegated to domain-specific applications. This method was easy to implement but problematic in terms of maintenance and robustness. With implementation of the SPAR repository came the development of a generic function to mint new ARKs. A growing proportion of new identifier assignment is now performed by this generic function.

Since its early stages, the ARK system at the BnF has been tuned regularly to become easier to maintain and configure, although technical issues still remain. To keep a robust system that can be trusted by end-users, we have to consider an increasing diversity of applications, the number of ARKs involved, and the flow of incoming requests.

3.6. Main lessons learnt about persistent identifier curation

To allow operational persistent identifier curation at a non-expert level, core questions have to be answered. With our eight-year hindsight, the key questions could boil down to this check-list:

- Who should be contacted in your institution when new kinds of objects are to be given persistent identifiers or when persistent-id aware applications are defined or revised?
- What are your identifiers identifying?
- Will your identifiers be re-assigned over the long-term or not?
- How much can the underlying content change over time? Can objects be deleted?
- Which services and subparts do you want to reveal, if any, so that end-users can cite a specific portion of the resource and/or a particular variant of that object?

4. Standards gap analysis

4.1. Machine-readable commitments

No identifier, regardless of scheme, can tell us if it will prove to be persistent into the future. The best "it" can do is to tell us (via its NMA) enough about itself, its resource, and resource provider to help us judge how and when to use it. The story it tells must be able to convey such things as provider support policies, expected changes to the resource (e.g., none, or corrections only), and the nature of the provider itself. A persistence promise is not black or white. Instead it is multi-dimensional, suggesting a breakdown into metadata elements.

Because we assume people searching for resources at scale will usefully want to filter based on persistence promise attributes, it will be necessary to support machine-readable commitments expressible via metadata. As was described earlier, the ARK inflection, "??", is designed to gain access to metadata statements about providers' persistence promises. Unfortunately, the ARK standard does not specify how to create machine-readable persistence promises. This section explores some of the areas that metadata should cover in such machine-readable commitments.

Support policies

Support policies and commitments vary between institutions, collections, and even between resources within a collection. For example, users often expect unchanging content behind durable links to *published* content, but they expect dynamic content behind durable links (persistent identifiers) to *advertised* content, such as a home page, curated database, or per-second updated stream of sensor data.

Setting expectations about this "content invariance" (or lack thereof), is critical, because audiences often avoid one kind and seek out the other kind, or vice versa, depending on the situation. Both are legitimate uses of persistent identifiers. Prior work at NLM (Byrnes, 2000) suggests at least four kinds of content invariance:

- *correctable*: Previously recorded content may be corrected (only) at any time.

- *dynamic*: Previously recorded content may be overwritten arbitrarily at any time, provided the resulting new content continues to match its metadata description. For example, the NLM homepage and the local weather page may both advertise very persistent identifiers for content that is completely overwritten from time to time.
- *unchanging*: Previously recorded content will not change, but encodings and markup may change during a format migration.
- *bitstream*: The bitstream representing previously recorded content will not change.

Datasets that grow

There is an important dimension of content invariance describing resources that grow, but whose growth pattern does not alter previously recorded content. We might describe such resources as subject to *non-disruptive growth,* as it is concerned with growth that does not in itself disrupt or displace previously recorded content. This applies to many common information resources, such as live, sensor-based data feeds, citation databases, and even serial publications.

The nature of the provider

Anyone can promise anything, but we might value a promise from one source more than from another. Relevant factors include not only what a provider promises in regard to identifier and resource support, but also how that provider is motivated, supported, and perceived. Thus mission, profit motive, succession plan, and reputation come to bear. Work to be done includes expressing these via metadata.

Support level

What are the provider's naming practices? How often is the collection inspected for broken identifiers? What action is taken when outages occur, and at what priority? Realistically, not all resources are equally important to a provider and its audience. To better support some resources means lowering priority support for other resources. What is a resource's "track record" and can one inspect it? These are all questions that can inform user choices of identifier.

4.2. Using ARKs in a semantic web context: investigating best practices

When the ARK specification came out in 2001, the core semantic web concepts and standards were already out or on their way (RDF was released in 1999). However, as the semantic web gained wider adoption, new best practices about URIs emerged over the next decade (W3C, 2008) and it is timely to re-evaluate the ARK specification in this new context. The main observation is that on one hand, ARKs can be embedded in URIs, which allows their use in the web of data, but on the other hand, the linked data best practices call for "Cool URIs" that, among other properties, "don't change" (Berners-Lee, 1998). For institutions that implement them, ARKs are a natural way to push identified resources onto the web of data. The question now is how to reconcile these two normative contexts at the BnF while implementing ARKs on the data.bnf.fr linked data service.

One could first ask how those two contexts address the question of multiple representations of a resource. On the semantic web, content negotiation using a generic URI yields the relevant representation of a resource; whether to reveal specific URIs for the variants is up to the content provider to decide. There is no reason why a provider could not implement an unqualified ARK name and rely on content negotiation to return linguistic or format variants to the user; or the user can reveal these variants by using traditional qualifiers[7].

[7] For the moment however, data.bnf.fr does not use ARK-URIs for its content negotiation. Early in the project when such choices were made, non-opaque URIs were considered better for SEO, as visibility on the web was one of the core aims of data.bnf.fr. Therefore, http://data.bnf.fr/ark:/12148/cb118905823 redirects to the temporary URI http://data.bnf.fr/11890582/charles_baudelaire/, which provides access to a particular representation of the object depending on the result of content negotiation (RDF/XML,

However, the real question is about the form of the URIs. In the early semantic web, a good deal of debate was about "real-world resources" that can be described on the web of data (with URIs), but could only be put *on* the web via substitutes (e.g. a description and/or a web page). It was initially considered wiser to use non-dereferenceable URIs. Non-HTTP URI schemes like "urn:" could be used to that end, and "info:" was explicitly defined for that purpose. By the end of the 2000's however, there was global consensus that an HTTP URI could be used for any resource. As a result, putting resources on the web of data now implies using HTTP URIs, i.e. URLs. This poses no conflict with ARKs since they are designed to be embedded in URLs using an NMAH that resolves them.

The main conflict between ARKs and URIs used on the semantic web concerns the qualifier part. At issue is distinguishing between a descriptive resource (available on a web page) and its underlying content (which might, or might not, be interpreted as a web page):

> "It is important to understand that using URIs, it is possible to identify both a thing (which may exist outside of the Web) and a Web document describing the thing. For example the person Alice is described on her homepage. Bob may not like the look of the homepage, but fancy the person Alice. So two URIs are needed, one for Alice, one for the homepage or a RDF document describing Alice. The question is where to draw the line between the case where either is possible and the case where only descriptions are available." (W3C, 2008).

With ARKs, the URI to reference the descriptive resource is constructed by adding the "?" inflection to the URI of the content resource. Unfortunately, supporting the single "?" (what looks like an empty query string) directly was impossible with the BnF infrastructure. What's more, BnF made the implementation choice to create ARKs directly for descriptive resource (e.g. authority records), so the mechanism needed was the opposite: *from* the identified descriptive resource (identified with an ARK name) *to* its underlying content resource, not the other way round. Therefore, we had to consider the other two mainstream choices:

- "**suffix hash URI**": you have http://example.com/resource for a web resource (e.g. a web page about a person), and http://example.com/resource#classifier for the underlying thing (e.g. the person itself). A browser client automatically strips off the # for consumption, which relies on standard web architecture and best practices.
- "**prefix slash URI**": you have http://example.com/doc/resource for the web document and http://example.com/id/resource for the underlying thing. This requires an HTTP 303 redirect from the resource URI to the URI of the web document.

The semantic web best practices highlight an area currently unaddressed by ARK qualifiers: how to name the underlying "thing" when the ARK is assigned to a descriptive resource. This is clearly not a whole-part problem (addressed by "/). Neither is it really a "service" or "variant" qualifier (addressed by ".") because the two identified things are quite distinct.

With ARKs only the "prefix slash URI" strategy is possible for the current state of the standard, which means using e.g. http://data.bnf.fr/**id**/ark:/12148/ark:/12148/cb118905823 (the French poet Charles Baudelaire) and http://data.bnf.fr/**doc**/ark:/12148/cb118905823 (the record describing him). This was not implemented because the redirection rules would present too great an extra server burden for our application.

From a technical standpoint, in data.bnf.fr the decision was made to locally extend ARKs and use "hash URIs". For example, we separate http://data.bnf.fr/ark:/12148/cb118905823 (web page about Charles Baudelaire) from http://data.bnf.fr/doc/ark:/12148/cb118905823#foaf:Person (Charles Baudelaire himself).

Notation3, N-Triples, JSON, or HTML, and language variants). We intend to reconsider this question with the evolution of SEO practices.

Looking back at the standard, would accommodating this change mean defining a new kind of qualifier, beginning with #, to name the underlying resource? Though technically possible, this would cause backwards compatibility issues, because the # character is not reserved in ARK names. In other terms, one could perfectly define the following (unqualified) ARK core identifier: ark:/9999/c5j3r4#hz45, with a # in the ARK name itself. Defining a # qualifier would break backwards compatibility in such cases. On the other hand, # already has a use in the standard web architecture (fragment for a URL) which makes it unlikely that implementers will use this character in their own implementation. A comprehensive survey of ARK implementers would be useful before any decision. If a # qualifier proved to be possible, we believe this would be a valid scenario to reconcile semantic web and ARK implementation approaches.

Conclusion

This article intended to look back at the history of using ARK persistent identifiers in one institution, and possible evolutions of the standard. Standards-wise, the question boils down to whether we should consider expanding the core features to increase cross-resolver interoperability and adapt ARKs to new contexts, or should we stick to the current ARK recommendation, which is flexible, simple, easy to use, and in most cases successful? Such questions will be taken up in follow-on work with the implementer community.

References

Archer, Phil. (2013) Study on persistent URIs: with identification of best practices and recommendations on the topic for the Member States and the European Commission. Retrieved May 02, 2014, from http://philarcher.org/diary/2013/uripersistence .

Bermès, Emmanuelle. (2006). Des identifiants pérennes pour les ressources numériques. Retrieved May 02, 2014, from http://2007.jres.org/planning/pdf/163.pdf.

Berners-Lee, Tim. (1998). Cool URIs don't change. Retrieved May 02, 2014, from http://www.w3.org/Provider/Style/URI.

BnF. (2013). URI and URL in data.bnf.fr. Retrieved May 02, 2014, from http://data.bnf.fr/en/semanticweb#Ancre3.

Byrnes, Margaret. (2000). Defining NLM's Commitment to the Permanence of Electronic Information. ARL 212:8-9. Retrieved May 07, 2014, from http://www.arl.org/newsltr/212/nlm.html

PREMIS Maintenance Activity. (2012). SPAR – Scalable Preservation and Archiving Repository; Retrieved May 02, 2014, from http://www.loc.gov/standards/premis/registry/premis-project_name.php?proj_ID=697.

CDL. (2013). ARK (Archival Resource Key) Identifiers. Retrieved May 02, 2014, from https://wiki.ucop.edu/display/Curation/ARK.

CDL. (2014). Registered Name Assigning Authority Numbers. Retrieved August 14, 2014, from http://www.cdlib.org/uc3/naan_table.html.

Hilse, Hans Werner, and Jochen Kothe. (2006). Implementing Persistent Identifiers. Consortium of European Research Libraries and European Commission on Preservation and Access. Retrieved May 02, 2014, from http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

IETF. (2013). The ARK Identifier Scheme. Internet-Draft. Retrieved May 02, 2014, from http://datatracker.ietf.org/doc/draft-kunze-ark.

Kunze, John. (2003). Towards Electronic Persistence Using ARK Identifiers. California Digital Library. Retrieved May 02, 2014, from https://wiki.ucop.edu/download/attachments/16744455/arkcdl.pdf.

Kunze, John and Adrian Turner. (2010). The ARK Identifier Scheme. Retrieved August 14, 2014, from http://dublincore.org/groups/kernel/spec/.

W3C. (2005). Uniform Resource Identifier (URI): Generic Syntax. Retrieved May 02, 2014, from http://www.ietf.org/rfc/rfc3986.txt.

W3C. (2008). Cool URIs for the Semantic Web. Retrieved May 02, 2014, from http://www.w3.org/TR/cooluris/.