ARK experts day

@ National Library of France (BnF)
March 22nd 2018

Participants:

- Emmanuelle Bermès (BnF) (points 4 & 5)
- Bertrand Caron (BnF) (all day)
- Angela Dappert (digital preservation consultant) (all day)
- John Deck (University of California, Berkeley) (points 1 & 2)
- Adrien Di Mascio (Logilab) (all day)
- Greg Janée (CDL) (points 1 & 2)
- Frédérique Joannic-Seta (BnF) (points 4 & 5)
- John Kunze (CDL) (all day)
- Amy Kirchoff (Portico) (all day)
- Thomas Ledoux (BnF) (all day)
- Roxana Maurer-Popistasu (Bibliothèque nationale de Luxembourg) (all day)
- Pascale Montmartin (Bibliothèque et Archives nationales du Québec) (points 3, 4 & 5)
- Sheila Morrissey (Portico) (all day)
- Sébastien Peyrard (BnF) (all day)
- Mark Phillips (University of North Texas) (all day)
- Claire Sibille de Grimoüard (Service interministériel des archives de France) (points 3, 4 & 5)
- Jean-Philippe Tramoni (BnF) (all day)
- Hélène Zettel (Service interministériel des archives de France) (all day)

1. ARK specification change proposals

(present IETF draft at https://tools.ietf.org/html/draft-kunze-ark-18)

a. Goals of changes to the spec
    i. fix what's broken, if anything
    ii. remove barriers to acceptance
    iii. move from Draft to RFC

5+1 proposed changes:

A. Literal character repertoire changes: allow '~', but disallow '#' (which is is reserved in URIs fragments and LOD).

- ‘~’ is filesystem friendly and will not cause problems in current situation
- The number of characters in the repertoire will not change except these 2 characters
- BnF: use the W3C recommendations for coolURIs: https://www.w3.org/TR/cooluris/ especially for hash URIs. Goal: distinguish the URI of a real-world object’s description from the URI for the RWO itself.
- Logilab is also in favour of disallowing the hash
- **No opinions against => the group approves the change**

B. Make the first ‘/’ optional, so that ark:/12345/678 is equivalent to ark:12345/678. This would match a near universal practice in other id schemes, and is a commonplace and understandable mistake that currently penalizes ARK users and potential adopters.

- Allows to reduce the length of the URL and users like the compactness of URLs
- It’s optional in the sense that the parsers should still accept it and not break any old ARKs
- Going forward, the canonical form would be the shorter form, but the documentation should always mention that the / is still accepted
- Would it make sense to accept ark:12148:cb12345689 → no
- The colons should be encoded according to the URI spec, but people are using it anyway -> colons are even a bigger problem
- **This is passed, with a note about the hierarchical computation**

C. Parsers (resolvers) should check for inflections (final punctuation character combinations) before normalization of final structural characters (‘/’ and ‘.’), for example, given “ark:/12345/678./”, parsers should check if “./” is an inflection and only normalize to “ark:/12345/678” if no inflection is matched

- When a resolver is checking an ARK in order to decide what to do, resolvers are sensitive to inflections and they normalize ARKs first and usually final . or / are left outside
- The idea is to keep the possibility to use those / and . for reserved used. E.g. direct consumption vs. landing page: would be great to leave the possibility open to use the final / for a landing page for the object
- When registering an ARK in a database, leave the inflections out and store the normalized form, but for resolving ARKs we should use these final characters and check if they are in the list of inflections
- For those final characters, instead of recommending to toss them out, recommend to “set them aside” because they might have a special meaning;

- A ? at the end of the id is not considered part of the identifier. Proposal to tell in the spec that reserved characters (with a list of inflections) at the end are not part of the identifier. This would be a set of rules for resolvers
- Some concerns about overloading the ARK specification with REST API type requests
- This change would not affect current ARKs. Inflections should only be used when you are using a resolver, checking what was at the end and whether it has an associated feature or not in the context
- Inflections are a situational use of the ARK
- A / at the end of an ARK is a violation of the current specification
- Software doesn't want a landing page, but users do want to make a choice
- People want access to previous versions or the history of the object; this kind of use would be possible with this change in the specification
- This is about implementing a resolver, and not so much about the ARK proper
- Proposal: leave new inflections out from the ARK spec, and elaborate the use cases in a separate specification that could also be used for other types of identifiers. Qualifiers would remain built in the ARK spec
- Question: should we leave all inflections out of the ARK spec and define them for all kinds of Identifiers? -> but very disruptive to the current spec.
- How do inflections and suffix pass-through work? Can we put an inflection at the end of a suffix? -> It should be subject to the same rules we were talking about, but not add this to the ARK specs
- **OK from the group, but will still look at the final wording**

D. Make the NAAN more flexible – instead of just 5 digits or 9 digits, allow any "beta-numeric" string (defined to be the same as noid repertoire: bcdfghjkmnpqrstvwxz0-9) with no runs of adjacent letters longer than two, eg, ark:/bc8/… but not ark:/bcd8/….
   - Leave lots of room for lots of name assigning authorities
   - Original idea: 5 digits can not be mistaken for a date, same for 9 digit.
   - People now want shorter addresses, so using a 9 digit NAAN would not be that desirable. So we might want to densify the NAAN namespace (allow letters in that mixture but make everything possible to disallow a brand)
   - Use the same opaque naming and character set as NOID, same rules (not more than 2 letters in a row).
   - Proposal: do not block this in the spec, but suggest very strongly absolutely no changes to the way we currently assign NAANs

- Have a separate policy that defines who gets which type of NAANs (NAANs other than the 5 digits the CDL currently assigns)
- Minimum length: the NAAN could be a single character. With this change we could also have NAANs for namespaces marked by special starting character (for example “x” for UUIDs or “p” for physical objects), not just organisations.
- With this change is it still a “Number”? We could keep the “NAAN”, but call it a “Name”.
- In terms of the registry indexing might be an issue if switching from integer to text
- The policy could be decided within the project “ARKs in the Open”
- **Everyone agrees with this change**

E. Update our understanding of what it means for metadata returned by inflections (‘?’ and ‘??’) in 2018 to be both human- and machine-readable. In 2003, a simple email-header format (eg, ANVL) served both purposes, but now it is common to see a human-readable HTML landing page with machine-readable metadata embedded in it (where it doesn’t interfere with the user experience).
   - Now we have new norms for this: enter this in a browser and get an HTML page that would be human readable, underneath: Javascript, JSON, Metadata tags, number of ways metadata can be embedded in HTML
   - The spec should recommend providing human- and machine-readable metadata, but not specifying its form (in particular, keep silent about content negotiation), just providing examples.
   - Resolvers would have a choice in what format they would return upon the inflections
   - Keep the resolving (including content negotiation and inflections) for a separate specification
   - All of these changes are subject to reviewing the final wording of the draft
   - Simple rewording: provide an example (“e.g., HTML page with embedded JSON-LD”...)
   - **OK with the group**

F. Max link length for the ARKs : now 128 digit limit
   - Should we raise it? Leave it alone
   - Whenever you are running a database, you have to set a limit for your column. However this is implementation-specific.
   - The idea is to remove obstacles, but the current trend is to have

    - Qualifiers are included in this limit (the BnF has already longer links when considering the qualifiers)
    - Current limits for databases shouldn’t be the reason for limiting something in the specification
    - 2 sentences would be struck from the spec and not mention the limit
    - Would dropping the limit have any impact on the suffix pass-through? -> No
    - The spec could just make a recommendation, but specific implementations could have a higher limit
    - **We will change the spec to eliminate the limit, but we will make a recommendation for a minimum of 255 characters supported in implementations**

2. Counting ARKs project

It is a feature of ARKs that there's no centralized maintenance authority, but that makes it difficult to count how many ARKs there are in the world. We propose an easy way for registered ARK implementers – those who are willing – to post a small JSON or YAML file (eg, at a well-known URL path) containing a date and an estimated number of ARKs published. Such files would be harvested to obtain a base total.

- Good for advocacy: how big is the number of ARKs do exist in the world? Not a one-time measure, but something that is updated on a regular basis
- Machine-readable assertion with a date and a number (very rough estimate of ARKs created at this date). Should be kept very simple.
- Updates: once a day or once a year - up to each institution.
- A bit complex for BnF, which manages several independent subnaming authorities - other institutions could be in this case.
- Maybe add semantics about ARKs that were deleted, ARKs that are not public
- Question: what we count are qualified or unqualified ARKs?
    - This should maybe be decided by the name assigning authority, because they should know this rough estimate
    - Maybe we need to differentiate between: all ARKs generated, all ARKs that might be seen in the wild/public, and ARKs on artifacts?
    - 4 options : ARKs that you assigned, ARKs that are in the wild, ARKs that the NAA considers to be an intellectual unit
    - It would be hard to give an estimate number with variants if each variant has to be counted as one ARK
    - DOIs are increasingly assigned as arbitrary levels of granularity

- We could also have links to each NAA’s naming policy
- We could also have something similar to Thor Project’s minting dashboard
- We all agree that variants form of the same object need not be counted separately; where it gets tricky is the hierarchy. Let’s think not in terms of a tree, but a graph.
- We want to provide numbers that are comparable to those using other identifiers
- We should leave out things when things would get wild with answers to dynamic queries (potential infinite number of possibilities)
- => strong encouragement to link to the naming policy
- We could start with a first version of the count by asking the institutions participating in the ARK experts day to send in their numbers and analyse that first and only afterwards ask the 500 other institutions: “Let’s get some numbers and define what the hell they mean and only after that decide which are the useful numbers”
- Question for the survey: what is the thing you are assigning an ARK to?
- We may even need some new terminology for what an ARK is, we don’t have language to distinguish between “mother ARK” from “child ARK” or between variants. It’s useful to distinguish between resolvable and citable, but both are really important.
- We should be raising awareness about naming policies that could inform future adopters on their own policies.
- What numbers are easy to produce ?
    - Counting ARK names, not include qualifiers
    - Counting ARK names that are kept / active / published / viable / ”in the wild”, “out there somehow”, “meant to persist”
- Do not add up the counts for bananas and apples: for different sub-naming authorities give separate numbers

3. Persistence statements

It has long been said that ARKs should provide a commitment or policy statement on demand from the current archival institution (object provider, name mapping authority). The day when this becomes true is closer with publication of “Persistence Statements: Describing Digital Stickiness” (https://datascience.codata.org/articles/10.5334/dsj-2017-039/). The paper proposes certain controlled vocabulary terms as building blocks for exactly this purpose. All that is lacking is to select, review, and revise the terms (which we can evolve ourselves in a crowdsourced metadata dictionary), and finally test and propose as a community consensus.

- Idea that the service provider (NMA) can provide its own persistence policy
- Persistence is not a binary things, there are many nuances of it
- Explain what is supposed to change in the content behind the ARK
    - “Frozen”: no bit will change: pretty unusual case
    - “Keeping”: bits will change but it will look identical: format migration, compression change
    - “Fixing”: we will fix errors if we come upon them: corrections to content.
    - Growing content: latest issue / version
    - The content will change: the homepage of an institution
- How long would you keep this object:
    - “Indefinite” commitment: we’re not committing to anything, we don’t know
    - “Lifetime”: it will be around as long as we are around
    - “Subinfinite”: we have successional arrangements with other organizations. Content will still be there even though we are not around anymore.
- What does the unqualified identifier lead you to by default? Landing page? Stable/Latest version?
- If something happens to this object, what is your priority for replacing it?
    - No particular priority
    - High priority: we will do whatever it costs to restore access to it
- Hyperlinked terms in the document are part of a vocabulary, but are still in a draft form. Controlled terms accessible, temporary definitions.
- Define persistence by the nature of the organisation: what are your missions? are you for profit, not for profit?
- It would help to get a group of motivated testers and sketch up a testing plan
    - Do we think this is a priority?
    - Is there another approach?
    - Is the vocabulary rich enough to describe your situation? Is it too elaborated?
- Vision of the future: machine-readable, but for the time being keep it in natural language (paragraphs), but with hyperlinks to the definitions so that people talk about the same things
    - A paragraph for each ARK, but with different paragraphs for “mother ARK” and “child ARK” (an ARK with qualifiers)
    - Round-trip test with prose, test it with some institutions, redefine the vocabulary and at the end “translate” it into machine-readable statements
- How do we set expectations upfront about things that could happen? -> Some terminology. This is different from things that happened: some provenance trail
- PREMIS proposes semantic units to express “significant properties” (what the repository considers to be preserved along migrations) and “preservation level” (bit-level preservation / functional preservation). Have been practically implemented at Danish Royal Library
- There are 2 types of persistence statements: specific for the institution or for the object -> Some of the vocabulary has to do with the nature of the institution
- The user: the recipient of the identifier to make an informed judgement about whether it trusts the service: does reality comply with the persistence policy? What premise can we have

- Such vocabularies could be used by all persistent identifier systems. Handle has a way of providing such a service
- The paper uses a crowd-sourced “metadictionary” (http://www.yamz.net/), where people can upvote or downvote a term; each term has an identifier
- How does this metadata dictionary manage multilingualism? Wikidata has several labels in different languages for each item.
- BnL will have persistence statements in three or four languages and also machine-readable ones.
- First stage: human-readable English statements. Each of our institutions tries to define them on one or two resource categories?
- Be prepared to find out that those terms are not working and need to be enhanced, in which case you can suggest changes in YAMZ.net, start new terms...
- Institutions doing the tests
    - Portico
    - CDL
    - BnF
    - BnL
    - UNT
- Each institution will do the exercise individually and send the first versions to John and only afterwards we will start working collaboratively
- **Deadline for the first version of the statements & first conference call: Tuesday, 22nd of May at 8:00 PDT / 11:00 EDT / 17:00 CEST**

4. Towards ARK sustainability

Any persistent identifier system maintained solely by one organization is vulnerable, and ARK is no exception. CDL is seeking guidance on sustainability of the ARK “infrastructure” and on building a coalition of organizations with shared responsibility and governance. The ARK infrastructure includes the specification, the NAAN registry, the arks-forum googlegroup, and the N2T.net resolver (code, admin scripts, and primary and secondary servers).

- We don’t want a persistent identifier system to depend on a single organisation
- Take the ARK infrastructure and make it shared (the specs and the registry of NAANs, then the perimeter of the pilot could extend to the N2T resolver, policies, monitoring, testing, source code)
- Formal discussions with DuraSpace
    - DuraSpace could help with promotion, awareness, membership

    - They could set up a non-profit association in the US (they have the legal means)
    - Would the fact that we're talking about an US-based association would be a problem for non-US organisations? -> According to the BnF, probably not. We could work with a memorandum of understanding or form a consortium (that works for IIPC or, in a similar form, for IIIF). A consortium does not have a legal entity though and having an organisation that can take care of the legal aspects, like DuraSpace, could help, by hiring a maintainer / developer for example.
    - Duraspace has a model where all the contributors can decide which projects they use their money for
- THOR has delivered a report on sustainability models for PID service providers - Angela will send out the document - it is not yet online (interim version online, but interim)
- Organization Identifier Project: https://figshare.com/articles/ORG_ID_WG_Governance_Principles_and_Recommendations/5402002. A Way Forward - Web | PDF
- DuraSpace summit in early April, good time to discuss a proposal and make an announcement
- Brainstorming about the expected benefits:
    - continuity plan for N2T
    - shared roadmap for community development, priorities, tools
    - creating a website for ARKs. The domain arks.org has been registered and currently redirects to N2T but should have content on its own, as a community-owned thing. Documentation, tools, registry of NAANs, map of users etc. Could have
    - funding opportunities
    - advocacy for ARK (Duraspace: "ambassador training" for people to advocate for their solution)
    - conferences & meetings around ARK

5. ARK survey: joint BnF-CDL proposal

BnF and CDL would like to feedback on a proposed online survey to get a better understanding of the different ARK implementations.

- The survey idea came after a discussion between BnF and CDL
- Would an online survey be helpful? If yes, what questions could we ask ARK users?
- Sample questions include:
    - Why did you choose ARKs?
    - For what kinds of objects?

    - Do you mint and/or resolve ARKs?
    - Are there challenges that you encountered using ARKs?
    - Would you consider contributing (at any level) to ARK community sustainability?
- The results of the community survey could be used as well for the IETF specs and showing there is support for this standard
- The survey could also include:
    - Some examples of resolvable ARKs
    - Examples of persistence statements, if they use them
    - Iteration of major ARK features to see what groups support (qualifiers, inflection, 'wear their identifiers')
- Beware of not having a survey that is too long!
- CDL is interested in designing the survey, BnF could take the lead
- Because the French speaking community is not always comfortable with English, we could have the survey in English and French, but reduce drastically the number of open questions
- How can we advertise for that survey? Which channels can we use to promote it?
    - Google group
    - Mailing lists
    - Twitter
- Ithaka could review the survey before sending it out
- We should create a first draft by the 22nd of May

6. Wrap-up

- How can we communicate with each other as a group?
    - Emails for the beginning

ACTIONS By May 22, 2018

- ACTION: John will propose wording changes for the ARK spec and send with diffs for review by May 2, 2018
- ACTION: Each institution will send a snapshot set of numbers and naming policy links for the Counting ARKs project
- ACTION: Portico, CDL, BnF, BnL, UNT, persistence statements test
- ACTION: John asks Duraspace for a debriefing call with BnF regarding ARK Summit outcomes
- ACTION: John asks Duraspace and CDL for input on the survey
- ACTION: Sebastien will talk to BnF survey department