Inflection change proposal

Problem: existing ARK inflections ? and ?? have not been adopted widely, for reasons that include

- It can be hard for servers to detect a terminal '?' as different from the absence of a query string. It is in fact impossible in Tomcat, and requires rewrite rules in Apache.
    - unlike '...??' (legal URL), '...?' is not a "legal" URL, so software libraries don't pass it through
- Although '?' is intuitive and language-agnostic, it can also be puzzling to some people.
- The metadata to be returned was only vaguely defined (mostly by example).
- The metadata syntax (ANVL) was non-standard and largely defined by example.

Requirements and Desiderata

Karen/Bertrand

1. At minimum, ?info must resolve to a human readable landing page, and *should* provide a *gateway* to machine-readable metadata
2. It is *strongly recommended* that meta tags with [something like] DC are implemented (I'm suggesting this since they are simple html, and all orgs should be able to do something with those)
3. Secondary to this, organizations are encouraged to use whatever data format[s] is appropriate in their context as the machine-readable data version of ?info, but *encourage* that organizations:

    a. utilize an established metadata standard (like DC) where possible
    b. utilize an established serialization for their metadata such as XML, JSON, or an RDF serialization such as JSON-LD or Turtle.
    c. express the document type via the "Content-Type:" HTTP header.
    d. utilize either content negotiation or queries in the form "&format=[json|xml]" property to deal with alternative formats.

    Karen: I added c) as a suggestion. I don't know if you want to indicate a preferred serialization/standard beyond this, or specify minimal metadata fields (the who, what, etc.), or keep it very loose. We could then provide examples that lay out different flavors that are acceptable – I would be willing to contribute an example.

John

1. Some continuity with past
    a. human-readable metadata returned
    b. machine-readable metadata returned
    c. including persistence statements
    d. who/what/when/where paradigm (ERC)
    e. THUMP-like request protocol -- ?info(X,Y) vs ?info&arg1=X&arg2=Y
2. Never RDF
    a. unfortunately, JSON-LD is RDF; see tweet https://twitter.com/justin_littman/status/1206944465027584001
    b. however, widely used schema.org borrows elements names from JSON-LD and uses them in meta tags, which aren't at risk of RDF complexity

Proposed solution discussions, in reverse chronological order

2019.12.15 Draft Inflection Spec: "?info"

The *info inflection* is a string, "?info", that may be added to an ARK before resolving it in order to request the return of human- and machine-readable metadata describing the identified object and the commitment made to it by its provider. A successful response returns metadata content as HTML intended for human consumption, along with embedded JSON intended for machine consumption. Future extensions are expected that will permit the request and return of alternate formats. Embedded HTML meta tags that repeat some of the metadata using schema.org element names are recommended because not all processors recognize JSON metadata. It is acceptable in the short term also to recognize the older "?" and "??" inflections and to treat them as synonymous with "?info", but their behavior may change in future versions of the ARK specification.

For the sake of discussion, we define some new terms. Resolution of a given ARK (or any URL) may be a multi-stage process starting with the *first resolver* hostname appearing in the URL form of the ARK when it is submitted for resolution. Examples are n2t.net and ark.bnf.fr. The first resolver may forward (HTTP redirect) to a second resolver, which may in turn forward to another, and so forth. The *content resolver* is the HTTP server that returns object content directly (ie, without forwarding). The *metadata resolver* is the HTTP server that, in response to the info inflection, returns metadata content directly. For a given ARK, the metadata resolver may be on a different host from the content resolver. (On the other hand, all three resolvers might also be on the same host.) For example, the N2T.net resolver stores a preservation copy of object metadata and can be configured on a per-ARK basis to respond to the info inflection directly or to forward it.

The object metadata returned in response to the info inflection depends not only on the object's immediate descriptive attributes but also on the object type and its place in a constituent cluster. For example, an ARK identifying a published article could have immediate attributes such as author (who), title (what), and date (when), but also, because it is a publication, additional core attributes such as publisher and length (number of pages). The article, one of eight in a particular issue of a journal, might also have multiple versions, in multiple formats, and might contain logical parts such as Abstract, Article, Appendices, and References. These represent its *constituent cluster*, which is a set of objects with which a given object has any of the following relationships: hasPart, isPartOf, isSiblingOf, HasFormat, hasVersion. For example, our article isPartOf a journal issue, hasPart References, and hasFormat (s) PDF and HTML. Because of its place in the cluster, the article's metadata should contain a link to the issue of which it is part. Link relationships within the constituent cluster exist independent of whether they are ARKs or whether they are ARKs that use the reserved '/' and '.' characters.

```
<script type="application/json">
{
"id_requested": "...",
"id_normalized": "...",
"id_surrogate": "...",   # (optional) different id for digital surrogate, implies requested id is about physical
object
"id_up1": "...",         # (optional) different id for 1st interesting landing page "above" this level
"id_up2": "...",         # (optional) different id for 2nd interesting landing page "above" the up1 level
"report": {
  "_comment": "The next 5 elements are for very broad cross-domain interoperation.",
  "who": "National Cancer Institute; ICPSR - Interuniversity Consortium for Political and Social Research",
  "what": "Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973–1977",
  "when": "1984-05-03",
  "where": "https://n2t.net/ark:/12345/x408001.v2",
  "how": "(:mtype data) Dataset",
  "thumbnail": "...",

  "_comment": "metatype-dependent core",
  ...,

  "persistence":  {
    "object": [ "indefinite", "standard" ],
    "content": [ "keeping", "waxing" ],
    "identifier": [ "single_use", "opaque", "intraversioned", "unbranded" ],
    "provider": [ "mission", "nonprofit" ]
 },
 "cite-as": "https://n2t.net/ark:/12345/x408001.v2",

 "_comment": "domain-dependent metadata",
 "name": "Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973–1977",
 "producer": "National Cancer Institute",
 "archive": "ICPSR - Interuniversity Consortium for Political and Social Research",
 "datePublished": "1984-05-03",
 "dateModified": "2015-08-06T11:20:58Z",
 "version": "v2"
}
</script>

<!-- why? because not everyone recognizes JSON script metadata -->
<meta name="DC.identifier" content="ark:/12345/x408001.v2" scheme="DCTERMS.URI"/>
<meta name="DC.title" content="Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973–
1977"/>
<meta name="DC.creator" content="National Cancer Institute"/>
<meta name="DC.publisher" content="ICPSR - Interuniversity Consortium for Political and Social Research"/>
<meta name="DC.date" content="1984-05-03" scheme="DCTERMS.W3CDTF"/>
<meta name="DC.type" content="Dataset"/>
```

2019.11.26 strawdog JSON

Returns HTML with
a) embedded GeoJSON, which allows foreign members from JSON-LD
  why? because of high integration with widespread tools, like google search and instant map integration is visually powerful
b) embedded HTML meta tags
  why? because not everyone is extracting JSON-LD tags
c) metadata elements formatted for human reading per provider preference

```
<script type="application/ld+json">
{
"@context": "http://schema.org",
"@type": "Dataset",
"@id": "https://n2t.net/ark:/12345/x408001.v2",

"who": "National Cancer Institute; ICPSR - Interuniversity Consortium for Political and Social Research",
"what": "Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973–1977",
"when": "1984-05-03",
"where": "https://n2t.net/ark:/12345/x408001.v2",
"how": "(:mtype data) Dataset",

"kids": [
  "https://n2t.net/ark:/12345/x408001.v2/file.xsl",
  "https://n2t.net/ark:/12345/x408001.v2/file.csv",
  "https://n2t.net/ark:/12345/x408001.v2/file.pdf"
],
"parent": "https://n2t.net/ark:/12345/x408001",
"cite-as": "https://n2t.net/ark:/12345/x408001.v2",
"stickiness": [
  "_see: https://datascience.codata.org/articles/10.5334/dsj-2017-039/",
  "indefinite", "keeping", "intraversioned", "standard", "NR", "OP"
],

"name": "Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973–1977",
"author": "National Cancer Institute",
"publisher": "ICPSR - Interuniversity Consortium for Political and Social Research",
"datePublished": "1984-05-03",
"dateModified": "2015-08-06T11:20:58Z",
"version": "v2",
"Description": "This dataset was produced as part of the Surveillance, Epidemiology, and End Results (SEER)
Program to monitor the incidence of cancer and cancer survival rates in the United States, thus carrying out the
mandates of the National Cancer Act. The SEER Program had several objectives: to estimate the annual cancer
incidence in the United States, to examine trends in cancer patient survival, to identify cancer etiologic
factors, and to monitor trends in the incidence of cancer in selected geographic areas with respect to
demographic and social characteristics..."}
</script>

<!-- why? because not everyone recognizes JSON script metadata -->
<meta name="DC.identifier" content="ark:/12345/x408001.v2" scheme="DCTERMS.URI"/>
<meta name="DC.title" content="Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973–
1977"/>
<meta name="DC.creator" content="National Cancer Institute"/>
<meta name="DC.publisher" content="ICPSR - Interuniversity Consortium for Political and Social Research"/>
<meta name="DC.date" content="1984-05-03" scheme="DCTERMS.W3CDTF"/>
<meta name="DC.type" content="Dataset"/>
```

2019.11.04 a different proposal for the new ?info inflection

Proposed: for any ARK *X*, *X*?info should lead to an HTML-formatted "landing" document (page) with metadata embedded as JSON-LD. The metadata, in human- and machine-readable form, includes

1. The ARK *X*
2. Descriptive metadata:
    a. who
    b. what
    c. when
    d. where
    e. how (metatype, similar to resourcetype)
    f. domain-specific elements (eg, publications vs physical samples vs vocabulary terms)
3. PIDs to first-level variants (versions, formats, change history) and components of *X*, if any
4. PID to the first (immediate) logical ancestor of *X*
    a. eg, if *X* is a PDF variant of a document object, this points to the logical object ARK listing *X* along with its sibling HTML and MSWord forms
5. PID to the **last** (root) logical ancestor of *X*
    a. eg, if *X* is a section of a chapter of a book, this points to the book logical object
6. Change history, if any
7. Licensing and accessibility information
8. How to cite, including "cite-as" header
9. Persistence statement

A great example to follow would be the A data citation roadmap for scholarly data repositories.

2019.09.16 proposal for a new, explicit word-based inflection: ?info

- ?info requests metadata
- ?info required, but spec continues to reserve '?' and '??' as optional synonyms

- ?info requests anvl/erc, but the spec permits (as always) alternate formats
    - continues to use THUMP conventions with parenthesized args
    - ?info equivalent to ?info()

This is a small adjustment to the spec that doesn't quite specify how to request alternate formats, but cracks open the door to work that we can complete, not in the spec, but in the AITO context. An example of that might be the THUMP request:
?info()as(application/json)

2019.08.05 more discussion of collapsing existing ? and ?? into just ??

2019.07.15 Proposed: suppress '?' inflection (let it be optional), leaving just the '??' inflection

- as before, '??' requests kernel elements plus any persistence statement
- '??' easier to implement than '?' (the latter being impossible to detect in Tomcat)
- '?' may be supported by older implementations (briefer record)
- ... or should '?' be made identical to '??' ?