The Archive and Package (arcp) URI scheme

Identifier
https://doi.org/10.5281/zenodo.1312582

Date created
2018-07-15

Submitted to
Workshop on Research Objects (RO2018)

Authors
Stian Soiland-Reyes <https://orcid.org/0000-0001-9842-9718>, eScience Lab, School of Computer Science, The University of Manchester, UK

Marcos Cáceres <https://marcosc.com/>, Mozilla Corporation

Abstract
The arcp URI scheme is introduced for location-independent identifiers to consume or reference hypermedia and linked data resources bundled inside a file archive, as well as to resolve archived resources within programmatic frameworks for Research Objects.

Background

Archive formats like BagIt [1] have been recognized as important for preservation and transferring of datasets and other digital resources [2]. More specific examples include COMBINE archives [3] for systems biology, CDF [4] for astronomy data, as well as the more general HDF5 [5] which is also used for meteorological data. For the purpose of this article an *archive* is a collection of data files with related metadata, typically packaged as a compressed file like *.zip* or *.tar.gz*.

One challenge with regards to embedding Linked Data in such archives is how to reliably generate and resolve internal URLs, for instance `<dataset13.zip>` may contain an RDF Turtle file `<metadata/description.ttl>` to describe the CSV file `<data/survey.csv>` — but in order to correctly reference that file it will either have to use a relative path `<../data/survey.csv>` or some pre-existing Web URL like `<http://example.com/dataset13/survey.csv>`.

The *Research Object Bundle* [6] format suggested re-using the app URI scheme for minting absolute URIs from relative paths of resources within a ZIP file. The app URL scheme [7] was originally intended for packaged web applications, where each application would get their own namespace like `<app://c6179148-3cde-4435-8e66-304453f89d59/>` with paths resolved from the corresponding application package ZIP file. However the app URL scheme did not progress further on the W3C Recommendation track, and this approach was abandoned in favour of the combination of Web App Manifest [8] and Service Workers [9]. Together these technologies reuse the http/https origin URL of a downloaded application manifest together with relative links, while also allowing a web application to work offline.

The Archive and Package (arcp) URI scheme

Inspired by the app URL scheme we defined the Archive and Package (arcp) URI scheme [10], an IETF Internet-Draft which specifies how to mint URIs to reference resources within any archive or package, independent of archive format or location.

The primary use case for *arcp* is for consuming applications, which may receive an archive through various ways, like file upload from a web browser or by reference to a dataset in a repository like Zenodo or FigShare.

In order to parse Linked Data resources (say to expose them for SPARQL queries), they will need to generate a *base URL* for the root of the archive.

It should be clear that using local file URIs [11] for extracted archives like `<file:///tmp/tmp.cUK6ERfdBe/>` do not serve well for this purpose, as they are not universally unique, are difficult to create consistently, and may introduce security risks of attacks like `<../../etc/passwd>`. Similarly it may be inappropriate to mint new web based URIs like `<http://repo.example.com/cUK6ERfdBe/>` as web presence should not be a requirement to process a linked data archive, in particular as processing may occur on a laptop or a cloud node with no public IP address.

Identifier structure

By definition an arcp identifier is an URI [12] with three parts:

```
<arcp://prefix,namespace/path>
```

Figure 1. Structure of arcp identifier

The arcp Internet-Draft specifies three initial *prefix* values: `uuid`, `ni` and `name`, each which defines how to identify a particular archive by a corresponding *namespace*. These namespaces are not intended to be directly resolvable without prior knowledge of the corresponding archive.

The *path* is the folder and file path within the archive, represented as an URI path [12] e.g. `/file.txt` or `/my%20project/about/intro.doc` — using percent-escaping where needed. The root folder `/` represent the archive itself.

UUID-based identifiers

The simplest case for temporary sandbox processing of an archive with arcp is to generate a new random UUIDv4 [13], e.g. `c6179148-3cde-4435-8e66-304453f89d59`, then the corresponding base URI is `<arcp://uuid,c6179148-3cde-4435-8e66-304453f89d59/>`, finding resources like `<arcp://uuid,c6179148-3cde-4435-8e66-304453f89d59/metadata/description.ttl>` referencing `<arcp://uuid,c6179148-3cde-4435-8e66-304453f89d59/data/survey.csv>`. The application is then able to do translation from arcp to local paths using URI parsing libraries to select the *URI path*, and augment that to the locally extracted path. Such arcp identifiers are temporary in nature, but the application can maintain a mapping from the UUID to the archive and perform extraction on demand, or the archive can self-declare its UUID, such as the `External-Identifier` header in BagIt [1].

arcp also suggests how a UUID can be reliably created from the URL location of an archive, thus if the application is processing `<http://example.com/download/archive13.zip>` it can use the name-based UUIDv5 [13] by SHA1 hashing the URL string to mint `<arcp://d9f0b57d-0504-5e9a-abae-f5f2b8c49b94/>` — with this method anyone processing that archive URL will always get the same arcp base URI, however the application will still need to maintain a mapping to find the original archive URL. Location-based arcp identifiers may also not be ideal for preservation purposes, as the archive might change upstream or move to a different location.

Hash-based identifiers

For this arcp defines a hash-based method, where the bytes of the archive file is used to find a checksum-based identifier based on the Naming Things With Hashes (ni) URI scheme [14]. For instance if the sha-256 checksum of a *zip* file is in hexadecimal `7f83b1657ff1fc53b92dc18148a1d65dfc2d4b1fa3d677284addd200126d9069` then the ni uri would be `<ni:///sha-256;f4OxZX_x_FO5LcGBSKHWXfwtSx-j1ncoSt3SABJtkGk>` by using the *base64* encoding of the checksum. The corresponding arcp base URIs for resources within the archive is then `<arcp://ni,sha-`

`256;f4OxZX_x_FO5LcGBSKHWXfwtSx-j1ncoSt3SABJtkGk/>`. With this method, anyone processing the byte-wise equal archive (using the same hash method) will get the same identifier.

Another advantage is that hash-identified archives can be retrieved from a NI resolver [14] using well known paths [15], e.g. `<http://repo.example.com/.well-known/ni/sha-256/f4OxZX_x_FO5LcGBSKHWXfwtSx-j1ncoSt3SABJtkGk>`. Clients can verify the checksum of the downloaded archive, so any resolver endpoint can be used.

Name-based identifiers

Finally, paying homage to its origin in app URLs, arcp can use a system-based app *name*. This is a suggested mechanism for resolving resources of an application package installed in a runtime system like Android applicationId or Java package name, where an application identifier can be directly reused in arcp for URIs within that runtime system, e.g. the URI `<arcp://name,com.example.myapp/styles/resource1.css>` references the resource `styles/resource1.css` within the installed package `com.example.myapp`.

As application package content do not necessarily correspond to archive file listings, it is open-ended how name-based arcp identifiers can be resolved, and indeed package content may vary per operating system, device type or application version, and so name-based arcp identifiers should be treated as system-local identifiers similar to `file:///` URIs [11], but within a particular programming framework.

Related work

Archive fragments

Without using arcp one could in theory still reference files within archives at an URL with fragments, e.g. `<http://example.com/download/archive13.zip#data/survey.csv>`, but most archive media formats like application/zip unfortunately do not define a fragment syntax, or are not even listed in the IANA media types registry (e.g. *tar.gz*), therefore this would be an ad-hoc approach which still needs to clarify details such as character escaping, if the root is `#` or `#/`, etc.

File URIs

As mentioned above, file URLs [11] representing local directories are fragile and not globally unique. It is perhaps less known that file URLs can have a host name, e.g. `<file://host.example.com/home/alice/extracted/archive13/>` (an empty hostname is equal to `localhost`). This approach, with a fully qualified domain name (FQDN), may be used if both the hostname and extracted path are stable, but this faces the same challenges as minting http/https URLs, which in many cases would be preferable as they are globally resolvable. An ad-hoc possibility here would be to use a UUID [13] as "host" to represent an archive's file system, technically permittable as the `file:` URL scheme [11] do not define any particular connection protocols, and an UUID is unlikely to be a valid hostname in DNS.

JAR URLs

If we restrict usage to ZIP files at a known URL, then they are in theory also valid *JAR files*, and we can address files with the jar URL scheme, e.g. `<jar:http://example.com/download/archive13.zip!/data/survey.csv>` — but here relative URIs may not parse well, as it is easy to accidentally climb out of `!/`, and technically the JAR URI scheme is missing the familiar `://` to indicate for URI parser libraries that it is indeed an hierarchical URI scheme [12].

Object Reuse and Exchange proxies

OAI-ORE [16] defines proxies to represent a resource as aggregated in a collection; these can be used to model archives [17], but ORE proxies face two problems: How to represent the file path, and how to identify the proxy so it can be used as a reference in Linked Data. The resource must be identified using two triples of `ore:proxyFor` (the archived file) and `ore:proxyIn` (the archive); but this reduces to the same problem of identifying the file. The ni URI [14] for the file bytes can in theory be used to identify the file, but the other missing information is the file path and name, which usually convey meaning for users.

The Research Object ontology's `FolderEntry` specializes the `ore:Proxy` to add a property `ro:entryName` to indicate the filename, but to find the full archive file path one would have to traverse the parent folder's `ro:entryName`. In either case there is no defined method to predictably generate unique identifiers for the ORE proxies themselves, although the RO Bundle specification recommend they should be randomly generated `urn:uuid` URIs, which would not be compatible with relative URIs within an archive.

```
@prefix ore: <http://www.openarchives.org/ore/terms/> .
@prefix ro: <http://purl.org/wf4ever/ro#> .

<urn:uuid:c5971b62-72e6-4a8f-8b0b-944065e0d5c8> a ore:Proxy, ro:FolderEntry ;
    ore:proxyFor <ni:///sha-256;f4OxZX_x_FO5LcGBSKHWXfwtSx-j1ncoSt3SABJtkGk> ;
    ore:proxyIn <urn:uuid:efb14c0a-3cd5-4d78-a168-f246d18bde39> ;
    ro:entryName "survey.csv" .
<urn:uuid:efb14c0a-3cd5-4d78-a168-f246d18bde39> a ore:Aggregation, ro:Folder .
<urn:uuid:24b34ecb-e46b-46ec-be36-a18dbba90247> a ore:Proxy, ro:FolderEntry ;
    ore:proxyFor <urn:uuid:efb14c0a-3cd5-4d78-a168-f246d18bde39> ;
    ore:proxyIn <http://example.com/download/archive13.zip> ;
    ro:entryName "data/" .
```

Figure 2. RDF Turtle example of how a file with the *sha256* checksum `7f83b1…6d9069` could be described using RO folders and ORE proxies to belong to `<data/survey.csv>` within the archive downloaded from `<http://example.com/download/archive13.zip>`

Publishing file systems as Linked Data

F2R [18], using the Nepomuk File Ontology [19], defines a way to publish file systems as Linked Data, where a server endpoint exposes the files and their file system metadata. F2R URIs are localized to an endpoint and an free-text named file system, e.g. `mysource`. Files are identified with UUIDs, e.g. `http://f2r.example.com/mysource/09b205be-bj80-4ab9-8ddc-802be95220bb`. Using the same example as for OAI-ORE we can combine F2R with PAV [20]:

```
@base <http://f2r.example.com/mysource/> .
@prefix nfo: <http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#> .

<c5971b62-72e6-4a8f-8b0b-944065e0d5c8> a nfo:ArchiveItem;
    nfo:fileName "survey.csv" ;
    nfo:belongsToContainer <24b34ecb-e46b-46ec-be36-a18dbba90247> .
<24b34ecb-e46b-46ec-be36-a18dbba90247> a nfo:ArchiveItem;
    nfo:fileName "data" ;
    nfo:belongsToContainer <5d0a538a-ef00-48b6-bcb2-f561effe9fe5> .
<5d0a538a-ef00-48b6-bcb2-f561effe9fe5> a nfo:ArchiveItem:
    nfo:fileName "archive13.zip" ;
    nfo:belongsToContainer <http://f2r.example.com/mysource/> ;
    pav:retrievedFrom <http://example.com/download/archive13.zip> .
<http://f2r.example.com/mysource/> a nfo:Filesystem .
```

Figure 3. RDF Turtle description of a file `<data/survey.csv>` within an archive `<http://example.com/download/archive13.zip>`, using Nepomuk File Ontology [19], PAV [20] and F2R [18] identifiers.

The F2R approach have similar disadvantages as JAR and OAI-ORE; in that the URIs do not support relative path resolution, that a web endpoint must be set up, and that the file paths are hidden through multiple steps. In addition one would need to assigned a corresponding file system name like `mysource`, although one may use a single file system as exemplified above and use `belongsToContainer` to treat archive files as if they are folders.

EPUB canonical fragment identifiers

EPUB is a standard for hypermedia eBooks. RO Bundle [6] is based on the EPUB Open Container Format [21]. EPUB Canonical Fragment Identifiers [22] can link to nested XML elements of an publication, for instance `<http://example.com/book.epub#epubcfi(/6/4[chap01ref]!/4[body01]/10[para05])>` use a variation of XPath with doubled indexes; here `/6` refer to the 3rd element of the root manifest's package element (which in ePub is always spine), then `/4[chap01ref]` is the second element itemref with XML id `chap01ref`, which reference is followed `!`, and then within the nested XML file `/4[body01]` is the 2nd element with id `body01`, traversed to find the 5th element called `para05`.

While this is quite a powerful construct that can refer to any XML element of nested documents, even sentences or words, it seems rather contrived and inflexible. The major limitation is that ePub archive resources are not identified by file paths, but must be addressable through rather rigid XML structures (order can't change), thus this approach is not appropriate for archives without an XML manifest. Even if using a RDF/XML manifest it would be inadvisable to assume a fixed order of it's XML elements. It seems however an appropriate reference scheme for ePub documents, as they have a fixed reading order.

arcp implementations

The **arcp Python library** [23] was developed to help creating, parsing and validating arcp URIs. In particular it can generate arcp based on random UUIDs, URL locations, names and hashing archive bytes. The arcp parser recognize the arcp prefix and can extract UUIDs or hashes, and can generate the corresponding `.well_known/ni` URI for retrieving the archive. This library is meant to complement Python's urlparse library, and so it is deemed out of scope for it to do any kind of resolution of arcp based on archive or network access.

The **Research Object Bundle library**, part of Apache Taverna (incubating), is adding support for arcp URIs in its opening and creation of RO bundles, initially using the arcp UUID format as a replacement for app URIs, with planned support also for hash-based identifiers and opening RO Bundles from a .well-known/ni endpoint.

The **CWLProv** [24] approach for capturing provenance of executing Common Workflow Language is using arcp in its BagIt External-Identifier to identify its research object.

```
External-Identifier: arcp://uuid,5d0a538a-ef00-48b6-bcb2-f561effe9fe5/
```

Figure 4. arcp as `External-Identifier` in `bag-info.txt` as declared in CWLProv.

For CWLProv the use of arcp is crucial, as it assigns global identifiers for use across resources in the RO bag, including the RO manifest itself and in W3C PROV file formats like PROV-N and N-Triples, neither which support relative URIs.

In this approach the UUID of the RO identifier `<arcp://uuid,82dee268-2411-45a2-83a9-3be14f84b754/>` also appears in the identifier `<urn:uuid:82dee268-2411-45a2-83a9-3be14f84b754>` of the top-level workflow run (the PROV Activity), and so this is showcasing how an RO that is the primary representation of a non-information resource (e.g. a process) can be identified using a directly derived arcp URI. While this could in theory also been achieved with an arcp UUIDv5 derived from the URL "location" of the activity `<urn:uuid:82dee268-2411-45a2-83a9-3be14f84b754>` that could be a confusing hack, as such URNs are not resolvable URLs. UUIDv5 hashing can however be appropriate for non-information resource that have a resolvable http/https permalink.

Conclusion

This article propose the arcp identifier scheme for resources within archives using formats like ZIP, tar and BagIt, and suggest arcp is useful for identifying standalone Research Objects and for processing Linked Data embedded in archives. The Internet-Draft draft-soilandreyes-arcp [10] is under consideration by IETF's Applications and Real-Time Area to progress towards Informational RFC status.

References

[1] J.A. Kunze, J. Littman, L. Madden, J. Scancella, C. Adams, The BagIt File Packaging Format (V1.0), Internet Engineering Task Force, 2018. https://datatracker.ietf.org/doc/html/draft-kunze-bagit-16

[2] Research Data Repository Interoperability WG, Research Data Repository Interoperability WG Final Recommendations, Research Data Alliance, 2018. https://doi.org/10.15497/RDA00025

[3] F.T. Bergmann, R. Adams, S. Moodie, J. Cooper, M. Glont, M. Golebiewski, et al., COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project., BMC Bioinformatics. 15 (2014) 369. https://doi.org/10.1186/s12859-014-0369-z

[4] Space Physics Data Facility, CDF Internal Format Description, 3.6, NASA / Goddard Space Flight Center, 2016. https://spdf.gsfc.nasa.gov/pub/software/cdf/doc/cdf364/cdf36ifd.pdf

[5] The HDF Group, HDF5 File Format Specification Version 3.0, The HDF Group, 2016. https://support.hdfgroup.org/HDF5/doc/H5.format.html

[6] S. Soiland-Reyes, M. Gamble, R. Haines, Research Object Bundle 1.0, researchobject.org, 2014. https://doi.org/10.5281/zenodo.12586

[7] System Applications Working Group, The app: URL Scheme, World Wide Web consortium, 2015. https://www.w3.org/TR/2015/NOTE-app-uri-20150723/

[8] M. Cáceres, K.R. Christiansen, M. Lamouri, A. Kostiainen, R. Dolin, M. Giuca, Web App Manifest, World Wide Web Consortium, 2018. https://www.w3.org/TR/2018/WD-appmanifest-20180704/

[9] A. Russel, J. Song, J. Archibald, M. Kruisselbrink, Service Workers 1, World Wide Web Consortium, 2017. https://www.w3.org/TR/2017/WD-service-workers-1-20171102/

[10] S. Soiland-Reyes, M. Cáceres, The Archive and Package (arcp) URI scheme, Internet-Draft. (2018). https://tools.ietf.org/html/draft-soilandreyes-arcp-03

[11] M. Kerwin, The “file” URI scheme, RFC Editor, 2017. https://doi.org/10.17487/RFC8089

[12] T. Berners-Lee, R. Fielding, L. Masinter, Uniform resource identifier (URI): generic syntax, RFC Editor, 2005. https://doi.org/10.17487/rfc3986

[13] P. Leach, M. Mealling, R. Salz, A universally unique identifier (UUID) URN namespace, RFC Editor, 2005. https://doi.org/10.17487/rfc4122

[14] S. Farrell, D. Kutscher, C. Dannewitz, B. Ohlman, A. Keranen, P. Hallam-Baker, Naming Things with Hashes, RFC Editor, 2013. https://doi.org/10.17487/rfc6920

[15] M. Nottingham, E. Hammer-Lahav, Defining Well-Known Uniform Resource Identifiers (URIs), RFC Editor, 2010. https://doi.org/10.17487/rfc5785

[16] C. Lynch, S. Parastatidis, N. Jacobs, H. Van de Sompel, C. Lagoze, The OAI-ORE effort: Progress, challenges, synergies, in: Proceedings of the 2007 Conference on Digital Libraries - JCDL ’07, ACM Press, New York, New York, USA, 2007: p. 80. https://doi.org/10.1145/1255175.1255190

[17] N. Ferro, G. Silvello, Modeling Archives by Means of OAI-ORE, in: M. Agosti, F. Esposito, S. Ferilli, N. Ferro (Eds.), Digital Libraries and Archives, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 216–227. https://doi.org/10.1007/978-3-642-35834-0_22

[18] Shaopeng He, Jianhui Li, Zhihong Shen, F2R: Publishing file systems as Linked Data, in: 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2013: pp. 767–772. https://10.1109/FSKD.2013.6816297

[19] Ansgar Bernardi, Gunnar Aastrand Grimnes, Tudor Groza, Simon Scerri, The NEPOMUK Semantic Desktop, in: Context and Semantics for Knowledge Management pp 255-273, 2011

[20] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A.J. Gray, C. Goble, T. Clark, PAV ontology: provenance, authoring and versioning., J. Biomed. Semantics. 4 (2013) 37. https://doi.org/10.1186/2041-1480-4-37

[21] EPUB Open Container Format (OCF) 3.1. W3C Member Submission 25 jan 2017. World Wide Web Consortium. https://www.w3.org/Submission/2017/SUBM-epub-ocf-20170125/

[22] EPUB Canonical Fragment Identifiers 1.1.Recommended Specification 5 January 2017. International Digital Publishing Forum. http://www.idpf.org/epub/linking/cfi/epub-cfi-20170105.html

[23] S. Soiland-Reyes, stain/arcp-py: arcp 0.2.0, Zenodo, 2018. https://doi.org/10.5281/zenodo.1165986

[24] F.Z. Khan, S. Soiland-Reyes, M.R. Crusoe, A. Lonie, R. Sinnott, CWLProv - Interoperable Retrospective Provenance capture and its challenges. Zenodo preprint, 2018. https://doi.org/10.5281/zenodo.1215611

Acknowledgements

This work has been done as part of the BioExcel CoE, a project funded by the European Union contract H2020-EINFRA-2015-1-675728.