

Lecture 21 – Identifiers

Advanced Data Management

Data Science & Scientific Computing / UniTS – DMG
Scientific and Data-intensive Computing / UniTS – DMG

- An identifier is a label which gives a name to an entity: a person, place, or thing.
 - Uniform Resource Locator
 - https://moodle2.units.it/pluginfile.php/221017/mod_resource/content/1/Lecture_07_XML.pdf
 - Fully qualified database indexed column
- Identifier is a common term
 - **Unique** IDs
 - Context / Namespacing
 - **Persistent** IDs
 - Someone guarantees its long lasting behaviour
 - 10 years? 1 century?



Uniqueness/Persistence



Passport: AA00000000 ← numero del documento
Tax Card: CODICE FISCALE **RSS MRA 70A41 F205Z**
Italian ID Card: REPUBBLICA ITALIANA, COMUNE DI, CARTA D'IDENTITA N° AZ 1234567
Social Security Card: SOCIAL SECURITY, THIS NUMBER HAS BEEN ESTABLISHED FOR JOHN DOE



RESETTA CREDENZIALI PER ACCEDERE AI SERVIZI IDEM
 This form will allow you to reset your password. A new password will be generated and sent to the email address on file for you.
 User ID:

 If you know your current password and want to change it, please visit the [password change](#) page.



ELECTRONIC TICKET BUSINESS 039
 MUSTERMANN/FRED MR
 FRA WUC LH 180 C 18AUG
 ZONE 2 7F 039
 MUSTERMANN/FRED MR
 ETKT 220 210025748-1
 FRA MUC
 LH 180 C 18AUG
 R25 2005 7F
 ZONE 2

Falls Sie diese Reiseinformation nicht oder nur teilweise lesen können, öffnen Sie bitte die angehängte PDF-Version.

Dieses ist eine automatisch erzeugte e-Mail. Bitte antworten Sie nicht

hierauf. **Lufthansa**

Buchungscode:
8RS1TB

Lufthansa Service Center

- The main developer of the ARK (Archival Resource Keys) system, John Kunze, has suggested that persistence simply means that
 - “an identifier is valid for **long enough**”
- Better reformulation:
 - persistent identifiers should only be assigned to resources that will be preserved for long term
 - that is, over several hardware and software generations
 - a persistent identifier and the services it provides should be at least as persistent as the resource identified
 - The resource may undergo several migrations and the outdated versions may no longer be accessible and/or usable
 - A user who has a persistent identifier of an old manifestation of a resource should be redirected to the latest version available, or to work level metadata, which may enable acquisition of the work in some other form, such as print

Persistent Identifiers in the Digital Era



- Actionability (resolve-ability)
 - In connection to digital objects and the internet
 - URLs → URIs → PIDs
 - ISBN “ISBN 951-45-9942-X” does not resolve to a work
 - But <http://urn.fi/URN:ISBN:951-45-9942-X> does
- Scope and Granularity
- Context
 - Permalink
 - URLs persistence
- Persistence requires management of resources identified
 - URIs can be used as persistent identifiers if they are properly managed
 - Domain Name Resolver / DNS

Persistent Identifiers & Services



- What an identifier resolves to
 - An object
 - A landing page including
 - Metadata
 - Links to actual resource
- Requires
 - Management
 - Translation services
 - Protocol?
 - Will HTTP(S) be always there?
 - schema/prefix
- It mainly is an organizational matter

Examples of Identifiers



- URN: Uniform Resource Name
- Handle(.net) System
- DOI: Digital Object Identifier
- ARK: Archival Resource Key
- PURL: Persistent URL (Uniform Resource Locator)
- URI: Uniform Resource Identifier
- UUID: Universally Unique ID
- ORCID: Open Research Contributor ID
- ADS bibcode: Astrophysics Data System bibliographic code
- IVOID: IVOA ID

Uniform Resource Name



- URN syntax

“urn:”<NID>”:”<NSS>

- where

- <NID> is a namespace identifier
 - to distinguish between different identifier schemes
- <NSS> is the namespace-specific string
- UTF-8 (UCS/2)

- example

- 'ISBN' is the NID for the ISBN
- ISBN URN example:
 - URN:ISBN:951-45-9942-X

- Not directly resolvable/actionable

- Can be used inside other PIDs systems: e.g. DOI.

- <http://dx.doi.org/10.1038/issn.1476-4687>

- Explicit statement for location preservation: proved unreliable

- Handles consist of a
 - prefix which identifies a "naming authority"
 - suffix which gives the "local name" of a resource
- Examples
 - 20.1000/100
 - 2381/12345
- Prefixes must be registered
- Handle is
 - Opaque: it encodes no information about the underlying resource
 - Provides only the means to retrieve metadata about the resource
 - UCS-2 based
- The Handle System is compatible with the Domain Name System (DNS)
 - but does not require it

Digital Object Identifiers



- DOI syntax

prefix/suffix

- 10.1002/joc.1130 is a valid DOI
 - better reported as doi:10.1002/joc.1130
- 10.1002. prefix composed of
 - 10: DOI identifier within the Handle system
 - 1002: identifier of the organization that has assigned the DOI
- joc.1130: suffix which identifies the resource
- DOIs are Handle persistent identifiers
- In practice, DOIs are usually expressed in the Web as hyperlinks:
 - <https://doi.org/10.1002/joc.1130>
 - Directly resolvable PID
 - DOI Foundation responsible for its dereference-ability
 - UTF-8

Archival Resource Key



- ARK syntax

[<http://NMAH/>]ark:/NAAN/Name[Qualifier]

- NAAN: Name Assigning Authority Number
 - mandatory unique identifier of the organization that originally named the object
- NMAH: Name Mapping Authority Host
 - optional and replaceable hostname of an organization that currently provides service for the object
- Qualifier: optional string that extends the base ARK to support access
 - to individual hierarchical subcomponents of an object (using a slash)
 - to variants (versions, languages, formats) of components (using dots)
- ARK example
 - <http://example.org/ark:/12025/654xz321/s3/f8.05v.tiff>
- Qualifier is the most prominent part of this PID
- ? and ?? added to retrieve metadata (brief) and preservation statement

Persistent Uniform Resource Locator



- PURL
 - uniform resource locator (URL, i.e., location-based URI)
 - used to redirect to the location of the requested web resource
 - PURLs redirect HTTP clients using HTTP status codes.
- The PURL concept is generic and can be used to designate any redirection service (named PURL resolver) given
 - a "root URL" as the resolver reference
 - <http://myPurlResolver.example>
 - "names" in the root URL
 - <http://myPurlResolver.example/name22>
 - means to provide
 - to associate each name with its URL to be redirected
 - to update this redirection-URL
 - persistence of the "root URL" and the PURL resolver

Uniform Resource Identifier (1)



- A URI provides a simple and extensible means for identifying a resource
 - Uniform
 - different types of resource identifiers can be used together
 - semantic interpretation of common syntactic conventions
 - introduction of new types of resource identifiers seamlessly
 - Resource
 - widest general sense
 - electronic document, image, source of information, service, collection of other resources
 - not necessarily accessible via the Internet: human beings, corporations, books
 - abstract concepts: operators, operands, relationship types, numeric values, ...
 - Identifier
 - distinguishing one resource from all other resources
 - an identifier does not define or embodies the identity of what is referenced
 - a system using URIs will not necessarily access the resource identified
 - "one" resource identified might not be singular in nature

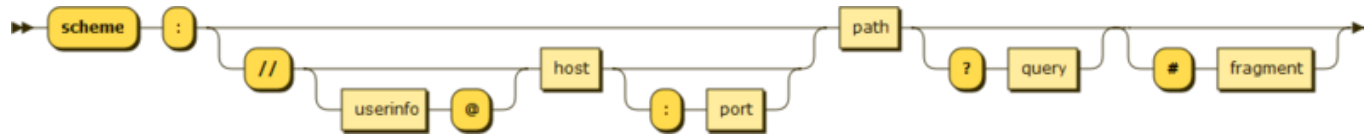
URI (2)



- URI Syntax

URI = scheme:[//authority]path[?query][#fragment]

authority = [userinfo@]host[:port]



```
      userinfo      host      port
      |             |             |
https://john.doe@www.example.com:123/forum/questions/?tag=networking&order=newest#top
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
scheme authority path query fragment

      userinfo      host      port
      |             |             |
ldap://[2001:db8::7]/c=GB?objectClass?one
|-----|-----|-----|-----|
scheme authority path query

mailto:John.Doe@example.com
|-----|-----|
scheme path

news:comp.infosystems.www.servers.unix
|-----|-----|
scheme path

tel:+1-816-555-1212
|-----|-----|
scheme path

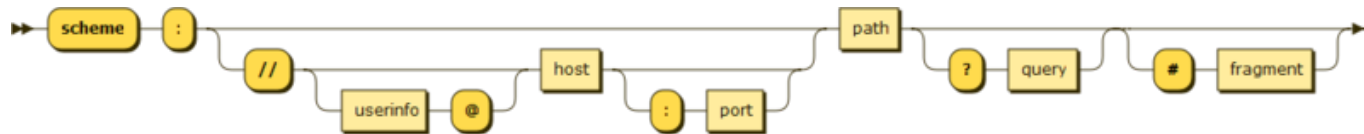
telnet://192.0.2.16:80/
|-----|-----|-----|
scheme authority path

urn:oasis:names:specification:docbook:dtd:xml:4.1.2
|-----|-----|-----|-----|
scheme path
```

URI (3)



- URI embeds
 - URL: explicitly serves a location to access the resource
 - URN: preserves the specification about “named” resources
- Involves identification/interaction distinction
 - Is used for identification
- URI is hierarchical
 - Starting from the “scheme” part down



```
reserved    = gen-delims / sub-delims
gen-delims  = ":" / "/" / "?" / "#" / "[" / "]" / "@"
sub-delims  = "!" / "$" / "&" / "'" / "(" / ")"
              / "*" / "+" / "," / ";" / "="
```

Universally Unique Identifier



- UUID is a 128-bit number used to identify information in computer systems
 - UUIDs are for practical purposes unique
 - uniqueness independent of a central registration authority or coordination
 - The probability that a UUID will be duplicated is not zero, it is close enough to zero to be negligible.
- Canonical textual representation
 - xxxxxxxx-xxxx-Mxxx-Nxxx-xxxxxxxxxxxx
 - the sixteen octets represented as 32 hexadecimal (base 16) digits
 - displayed in five groups separated by hyphens [8-4-4-4-12]
 - Example: 123e4567-e89b-12d3-a456-426655440000
 - M indicate the UUID version, N indicate the UUID variant
 - i.e. a way to get the meaning of the numbers, if needed

Open Research Contributor Identifier



- ORCID
 - an https URI with a 16-digit number
 - e.g. <https://orcid.org/0000-0001-2345-6789>
 - ISNI compatible / URN valid
 - ISNI: International Standard Name Identifier
 - Random generated 16 digits
- Identifies a person
 - Uniqueness in author/paper cross-reference
 - Not fail-proof but should really be unique
- Attaches to a record of that person

- Astrophysics Data System (abstracts service) bibliographic codes
 - YYYYJJJJVVVVMPPPPA – always 19 characters, “.” padded
 - YYYY: year of publication
 - JJJJ: standard abbreviation for the journal (left padded)
 - ApJ, AJ, MNRAS, Sci, PASP, ... internally controlled
 - VVVV: volume number (for a serial) or a type abbreviation (right padded)
 - conf, meet, book, coll, proc
 - M: Qualifier for publication:
 - E: Electronic Abstract (usually a counter, not a page number)
 - L: Letter
 - P: Pink page
 - Q-Z: Unduplicating character for identical codes
 - PPPP: Page number (right padded)
 - Note that for page numbers greater than 9999, the page number is continued in the m column.
 - A: The first letter of the last name of the first author
 - 1992ApJ...400L...1W
 - Astrophysical Journal Letters volume 400, page L1 (Windhorst).
 - Resolver: <http://adsabs.harvard.edu/abs/1992ApJ...400L...1W>

IVOA Identifier (1)



- An IVOA identifier, or IVOID, is a special sort of URI
 - “ivo” as the scheme
 - “Registry part” including authority and path
 - “local part” including query and fragment
- `ivo://<authority><path>?<query>#<fragment>`
 - Registry part → `ivo://<authority><path>`
 - Registry reference, it must be resolvable in the Registry
 - Resource key → `path`
 - Authority → `ivo://<authority>`
 - Local part → `?query#fragment`
- IVOID example: `ivo://example.org/svc?voc.xml#Term`

IVOID (2)



- Restrictions to URI specification apply to
 - Authority
 - it MUST be at least three characters long
 - it MUST begin with an alpha-numeric character
 - it MUST NOT contain percent-encoded characters
 - it MUST NOT contain characters outside of <unreserved>, with the tilde strongly discouraged
 - there are no <userinfo> or <port> components
 - Resource key
 - No percent encoding
 - No sub-delimiters, unless explicitly specified by other IVOA REC
- Query & fragments follows URI
 - Including semantic meaning of fragments

`ivo://nasa.heasarc/user/STScI_1/1a-7z.u` → OK

`ivo://a2/data!g-vo.org` → KO (but...)

- Resolving IVOIDs

- It is done through a Registry query (bootstrap: you need the Registry endpoint)

- Registry references → resolve directly

- IVOID with query part

- Resolve the Registry reference (strip from ? on)

- Identify the service dereferencing the full IVOID

- Depends on resource context

- IVOID with fragment

- Same as IVOID up to the query part

- Fragment specification depends on the specific resource

- i.e. it's up to the service/application

↑
[F]indability

- IVOIDs comparison (restriction w.r.t. URI)

- As no hierarchy is implied in any IVOID part, no path segment normalization

- As IVOIDs must not percent-encode characters that do not need to be encoded, no percent-encoding normalization is ever performed on IVOIDs

- the resource key is also compared case-insensitively

- Same rules as IVOID, i.e. URI with ivo:// scheme and restrictions
 - Dataset Identifiers
 - DIDs / PubDIDs / CreatorDIDs
 - Answering the need to reference datasets
 - They are IVOIDs where the query part identifies (locally) the dataset
 - No need to register them
 - No need to have a service resolving them
 - Fragment part can be used to identify sub-parts
 - Standard Identifiers
 - To identify the IVOA endorsed standards
 - And use in annotating resources
 - Fall inside the ivo://ivoa.net authority
 - Have resource key and local part as
 - /std/<standard-ref> "#"<key-name> "-"<version>
 - e.g. ivo://ivoa.net/std/exampleProto#model-1.0

Goals for Identifiers



- Resource “thing” identification
 - Scientific work
 - Research data
 - Involved people
 - ...
- Allow reference among identified objects
 - Relationships can be identified too
- Usefulness if interaction/resolving is needed
 - Persistence of the identifier
 - Flexibility on the location and protocol
- Allow granularity if needed

RDA Consolidated Assertions on PIDs



- PIDs are increasingly important and are being applied almost everywhere across *sectors* and disciplines, and for all types of digital objects. (Here, the term "*sectors*" covers science, industry, governments, health care, etc.)
- Data management experts are becoming increasingly dependent on the availability of functioning persistent identifiers which
 - are uniquely identifying a specific Digital Object
 - in general consist of a name space indicator (prefix) and a local identifier suffix)
 - are actionable on the web, by extending it to a fully defined URI, if required
 - can be persistently resolved to state information and/or a landing page
 - are associated with a persistent resolution system
 - are issued and managed by a clearly specified registration authority

● Data Citation of Evolving Data

● Recommendations of the WG on Data Citation - RDA

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

A. *Preparing the Data and the Query Store*

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store Facilities:** Provide means for storing queries and the associated metadata in order to re-execute them in the future.

C. *Resolving PIDs and Retrieving the Data*

- **R11 – Landing Page:** Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet.
- **R12 – Machine Actionability:** Provide an API / machine actionable landing page to access metadata and data via query re-execution.

D. *Upon Modifications to the Data Infrastructure*

- **R13 – Technology Migration:** When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated fixity information.
- **R14 – Migration Verification:** Verify successful data and query migration, ensuring that queries can be re-executed correctly.

B. *Persistently Identify Specific Data Sets*

When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness:** Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.
- **R5 – Stable Sorting:** Ensure that the sorting of the records in the data set is unambiguous and reproducible
- **R6 – Result Set Verification:** Compute fixity information (checksum) of the query result set to enable verification of the correctness of a result upon re-execution.
- **R7 – Query Timestamping:** Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at the time a user issued a query.
- **R8 – Query PID:** Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID.
- **R9 – Store Query:** Store query and metadata (e.g. PID, original and normalized query, query & result set checksum, timestamp, superset PID, data set description, and other) in the query store.
- **R10 – Automated Citation Texts:** Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing the data. Include the PID into the citation text snippet.



June 26, 2018

Dataset Open Access

VAMDC extraction with identifier = 5c91e7c7-e0c9-474f-b76a-62bff7b38468

VAMDC, Consortium

This is a dataset extracted from <http://vamdc.icb.cnrs.fr/gecasda/tap/> VAMDC node.
Query originating this dataset: `query=select * where (inchikey in ['quzpnffhzprkjd-aklpvkdbsa-n', 'quzpnffhzprkjd-bjudxgmsa-n', 'quzpnffhzprkjd-igmarmgpsa-n', 'quzpnffhzprkjd-oiobtwansa-n', 'quzpnffhzprkjd-oubtvsysa-n']);`
Data source version: 1
Data format: XSAMS 12.07
Query uuid in VAMDC query store: 5c91e7c7-e0c9-474f-b76a-62bff7b38468

1 views 0 downloads

[See more details...](#)

Indexed in

Publication date:
June 26, 2018

DOI:
[DOI 10.5281/zenodo.1620783](https://doi.org/10.5281/zenodo.1620783)

Related identifiers:
Identical to:
<https://cite.vamdc.eu/references.html?uuid=5c91e7c7-e0c9-474f-b76a-62bff7b38468>

License (for files):
[Creative Commons Zero v1.0 Universal](#)

Preview

[5c91e7c7-e0c9-474f-b76a-62bff7b38468.zip](#)

[5c91e7c7-e0c9-474f-b76a-62bff7b38468.xsams](#) 133.6 MB

Files (2.1 MB)

Name	Size	
5c91e7c7-e0c9-474f-b76a-62bff7b38468.zip	2.1 MB	Preview Download

Versions

Version 1 10.5281/zenodo.1620783 Jun 26, 2018

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.1620782](https://doi.org/10.5281/zenodo.1620782). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

DOI [10.5281/zenodo.1620783](https://doi.org/10.5281/zenodo.1620783)

Data source : <http://vamdc.icb.cnrs.fr/gecasda/tap/>

Data source version : 1

Query : select * where (inchikey in ['quzpnffhzprkjd-aklpvkdbsa-n', 'quzpnffhzprkjd-bjudxgmsa-n', 'quzpnffhzprkjd-igmarmgpsa-n', 'quzpnffhzprkjd-oiobtwansa-n', 'quzpnffhzprkjd-oubtzvsysa-n'])

Query identifier : 5c91e7c7-e0c9-474f-b76a-62bff7b38468

Query result : [XSAMS file](#)(if not available, please try again in a few minutes)

XSAMS version : 12.07

Query result downloaded on (UTC+1) :

- 2018-11-23 16:31:56
- 2018-7-24 16:34:57
- 2018-7-13 16:25:08
- 2018-7-13 16:24:18
- 2018-7-13 16:20:24

References

- **Title** : Line positions and intensities for the ν_3 band of 5 isotopologues of germane for planetary applications
- **Journal** : Journal of Quantitative Spectroscopy and Radiative Transfer
- **Authors** : Boudon,V. and Grigoryan,T. and Philipot,F and Richard,C and Kwabia Tchana,F and Manceron,L. and Rizopoulos,A and Vander Auwera,J and Encrenaz,T
- **Pages** : 174,183
- **Volume** : 205
- **Year** : 2018
- **Reference name in bibtex** : BICB-GEH4-1

ADS data linking & ORCID



https://ui.adsabs.harvard.edu

astrophysics data system

Classic Form **Modern Form** Paper Form

QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms

author author:"huchra, john" citations citations(author:"huchra, j")

first author author:"^huchra, john" references references(author:"huchra, j")

abstract + title abs:"dark energy" reviews reviews("gamma-ray bursts")

year year:2000

year range year:2000-2005 refereed property:refereed

full text full:"gravitational waves" astronomy database:astronomy

publication bibstem:ApJ OR abs:(planet OR star)

Use a classic ADS-style form

Learn more about searching the ADS

Access ADS data with our API



- IVOID in Registry
 - Authority
 - datasets/collections
 - services/protocols
- StandardIDs
- Publisher DIDs