Use of DataCite DOIs for Citing Astronomical Data

ADASS XXVII BoF

2017-10-24

Arnold Rots, Raffaele D'Abrusco, Sherry Winkelman, Josh Peek

Context

- The Chandra Data Archive (CDA) has been linking its datasets to the literature since 1999, using *ivo* Persistent Identifiers (PID) under authority of the ADS since they were standardized by the ADEC
- This was a stopgap measure since at the time no globally accepted standard PID was available
- At this time it is pretty clear that DOIs registered at DataCite with metadata conforming with the DataCite metadata schema is the way to go
- CDA will transition to DataCite DOIs and MAST is in the process

Benefits

- Stable registry of PIDs
- Outsourced registry maintenance
- Global visibility and access
- Good metadata set
- Enhanced discoverability, also outside the astronomical search engines and portals

Issues to Consider

- The architectural design to be used for the DOIs in order to allow individual datasets to be identified, while also creating user-friendly landing sites from the literature
- The schema of the metadata record for dataset DOIs – primarily to optimize discoverability: guidance, recommendations, standards?
- Definition and use of trailing fragments – to allow access to individual components in the datasets

DataCite Metadata Schema

- **http://schema.datacite.org/meta/kernel-4.1/**
- Record builder (not perfect, but very helpful): **http://dspace.ut.ee/bitstream/handle/10062/53326/datacite_metadata_generator4.html?sequence=6&isAllowed=y**
- Publishers, Creators, Contributors, etc.
  - Can be persons or institutions
  - Can be named or indicated by an identifier
  - It would be desirable if one could standardize on ORCIDs
  - There is an effort underway to create ORCIDs for institutions – which would be very helpful!

Example

- The next slides step though the most relevant metadata elements, using an example from the CDA (ObsId 2000) to provide a sense of what might be possible
- It is based on version 4.0 of the DataCite metadata schema
- We try to take maximum advantage of the schema
- Colored items will be further discussed

Mandated Elements

- 1 Identifier: 10.5072/006.01.2000 (DOI)
  - Identifier type: “DOI”
- 2 Creator
  - Creator name: “CXC-DS”
  - Name identifier: <CXC-DS ORCID>
    - Name identifier scheme: “ORCID”
  - Affiliation: “Smithsonian Astrophysical Observatory”
- 3 Title: “Chandra X-ray Observatory ObsId 2000”

Mandated Elements - 2

- 4 Publisher: “Chandra X-ray Center/SAO”
- 5 Publication year: “2002” (year made publicly available)
- 10 Resource type: “High Energy Astrophysics Data/X-ray”
  - Resource type general: “Dataset”

Recommended Elements

- 6 Subject: “X-ray Data” or “Coordinates Chandra/HST observations of the Crab nebula” or <Unified Thesaurus>?

Recommended Elements - 2

- 7 Contributor:
  - Contributor type: “RightsHolder”
    - Contributor name: “NASA”
  - Contributor type: “HostingInstitution”
    - Contributor name: “SAO”
  - Contributor type: “DataManager”
    - Contributor name: “Chandra Data Archive”
  - Contributor type: “RegistrationAgency”
    - Contributor name: “Smithsonian Institution”
  - Contributor type: “Distributor”
    - Contributor name: “Chandra Data Archive”

Recommended Elements - 3

- 8 Date
  - “2001-04-11”
    - Date type: “Created”
  - “2002-04-11”
    - Date type: “Available”
  - “2004-03-17”
    - Date type: “Updated”
  - “2007-01-04”
    - Date type: “Updated”
  - “2012-09-18”
    - Date type: “Updated”

Recommended Elements - 4

- 12 Related identifier
  - “2002ApJ...577L..49H”
    - Related identifier type: “bibcode”
    - Relation type: “IsCitedBy”
  - “10.1086/344132”
    - Related identifier type: “DOI”
    - Relation type: “IsCitedBy”
  - “10.5072/006.02/1504”
    - Related identifier type: “DOI”
    - Relation type: “IsPartOf”

Recommended Elements - 5

- 17 Description
  - “Coordinated Chandra/HST Observations of the Crab Nebula”
  - Description type: “Abstract”
- 19 Funding reference
  - Funder name: “NASA”
  - Award title: “Chandra X-ray Center”
  - Award number: “NAS 8-03060”

Recommended Elements - 6

- 18 GeoLocation
  - GeoLocation place: “ICRS”
  - GeoLocation point
    - Point longitude: “83.632”
    - Point latitude: “22.016”
  - GeoLocation polygon: <FOV>
    - Polygon point
      - Point longitude: <>
      - Point latitude: <>
    - …

Optional Elements

- 11 Alternate identifier: “ivo://ADS/Sa.cxo/obsid/2000”
  - Alternate identifier type: “ADEC/ITWG”
- 13 Size:
  - “300 MB Primary”
  - “350 MB Secondary”
  - “2.64 ks Exposure”
- 14 Format: “FITS”
- 15 Version ?
- 16 Rights ?

Potential Issues

- There are three metadata concepts that have fuzzy boundaries for astronomical data archives:
  - 2 Creator: “CXC-DS”
  - 4 Publisher: “Chandra X-ray Center/SAO”
  - 7 Contributor: (several types)
  - Specific guidance on who-is-what is desirable

Potential Issues - 2

- 10 Resource type: “High Energy Astrophysics Data/X-ray”
- 6 Subject: “X-ray Data” or “Coordinates Chandra/HST observations of the Crab nebula” or <Unified Astronomical Thesaurus terms>? If UAT, the subject scheme needs to be provided

- Should one of these (Resource Type?) be used to identify all astronomical datasets as such?

Potential Issues - 3

- 13 Size:
  - "300 MB Primary"
  - "350 MB Secondary"
  - "2.64 ks Exposure"
- 15 Version ? Might be better taken care of by Dates and/or Related Identifier
- 16 Rights ? Should this be used for copyright? Doesn't look that's intended
- 18 GeoLocation
  - GeoLocation point: "ICRS"
  - We have proposed to DataCite to allow celestial coordinates

Additional Contributor Types

- ContactPerson
- DataCollector
- DataCurator
- Editor
- Producer
- ProjectLeader
- ProjectManager
- ProjectMember
- RegistrationAuthority
- RelatedPerson
- Researcher
- ResearchGroup
- Sponsor
- Supervisor
- WorkPackageLeader
- Other

Additional Relation Types

- Cites
- IsSupplementTo
- IsSupplementedBy
- IsContinuedBy
- Continues
- HasMetadata
- IsMetadataFor
- IsNewVersionOf
- IsPreviousVersionOf
- HasPart
- IsReferencedBy
- References
- IsDocumentedBy
- Documents
- IsCompiledBy
- Compiles
- IsVariantFormOf
- IsOriginalFormOf
- IsIdenticalTo
- IsReviewedBy
- Reviews
- IsDerivedFrom
- IsSourceOf

Additional …

- Date Type
  - Accepted
  - Copyrighted
  - Collected
  - Issued
  - Submitted
  - Valid
- Resource Type General
  - Audiovisual
  - Collection – for aggregating DOIs
  - Event
  - Image
  - InteractiveResource
  - Model
  - PhysicalObject
  - Service
  - Software
  - Sound
  - Text
  - Workflow
  - Other

Related Applications

- The Related Identifiers group of metadata elements (Relation Types) provides fairly rich support for provenance information; it would be desirable for IVOA Provenance recommendations to be consistent

- Version 4.1 of the DataCite metadata scheme includes enhanced support for software citation

More from the BoF Discussion

- Josh Peek described MAST initiative: aggregating DOI created at paper submission, pilot for STScI staff only
- Minters remain responsible for, and retain control over, updates to DOI metadata
- Be mindful of desirable personal credits
- Should this be considered for a IVOA standard?
- Don't lose sight of radio (and other ground-based) data

More from the BoF Discussion

- Acknowledgment recommendations
- The contents of the DataCite repository are not a replacement for our bibliographic databases; if anything, a partial mirror
- Scholix initiative (cf. RDA) aims to provide a registry of all PID connections – might be a resource for aggregations