Update on the DOI initiative at the Chandra Data Archive

Raffaele D'Abrusco

and

Arnold Rots, Sherry Winkelman
and the Archive Operations team

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Chandra interests

- The **Chandra Data Archive** has built a full bibliography for the mission, containing all articles using Chandra data, with full high-granularity linking to the observations
  - research tool
  - metrics of scientific impact of the observatory

- CDA has used/is using a provisional Persistent ID specification
  - agreed upon ~17 years ago by NASA data centers (ADEC) and the ADS
  - journals and data archives are pushing towards adoption of DOIs across the board for datasets

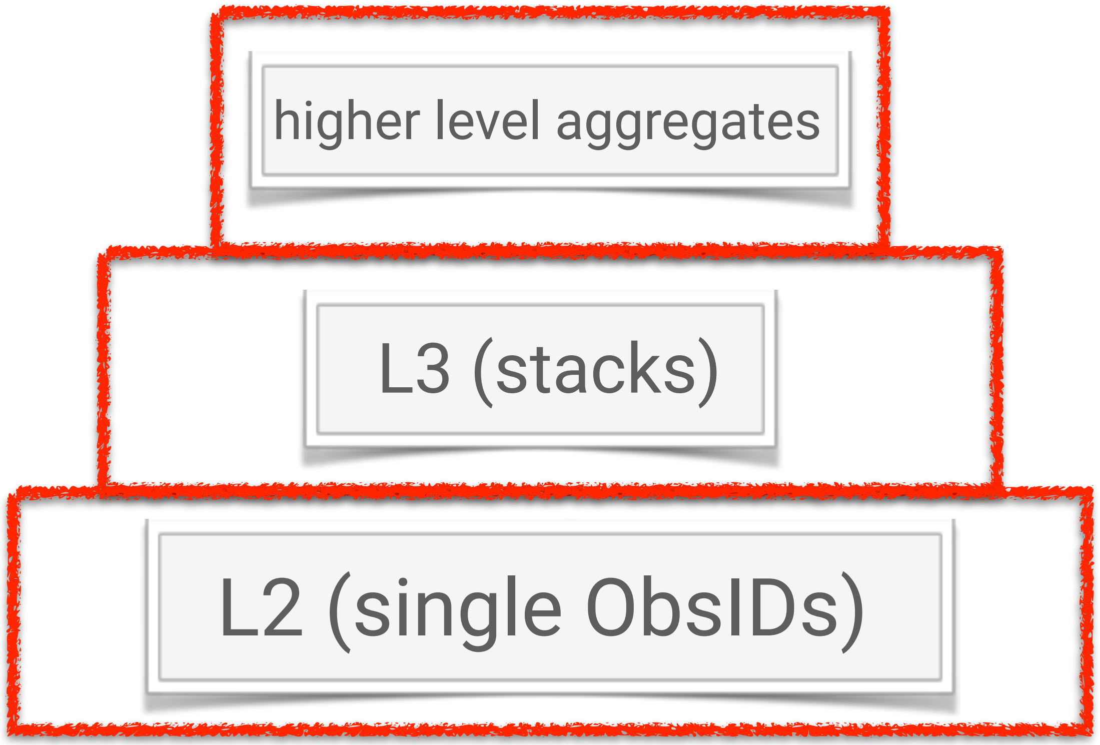
- We are working on the migration to **DataCite DOIs**

Why giving data objects identifiers?

- ➡ Three very good reasons:
  - ➡ provide scientists a tool to credit the data provider
  - ➡ document accurately what data was used to produce specific scientific results
  - ➡ provide enduring access to the data objects
- ➡ To achieve this, we need to:
  - ➡ label data objects with PIDs
  - ➡ encourage or enforce insertion of the PIDs in the manuscript
  - ➡ keep up-to-date record of connections/relations among different types of identifiers
  - ➡ follow the historical evolution of the archive
    - ➡ growth of the usage and importance of “advanced data products”

Something else

- ➡ We need a **formal and abstract description of the internal structure** of the Chandra data archive holdings
  - ➡ represent different **levels** and **types of aggregation** of datasets
  - ➡ provide visibility to “value-added”, merged datasets to increase scientific return
  - ➡ leave a blueprint of the complexity of the archive (and its growth over time) as a part of the scientific legacy of the mission



higher level aggregates

L3 (stacks)

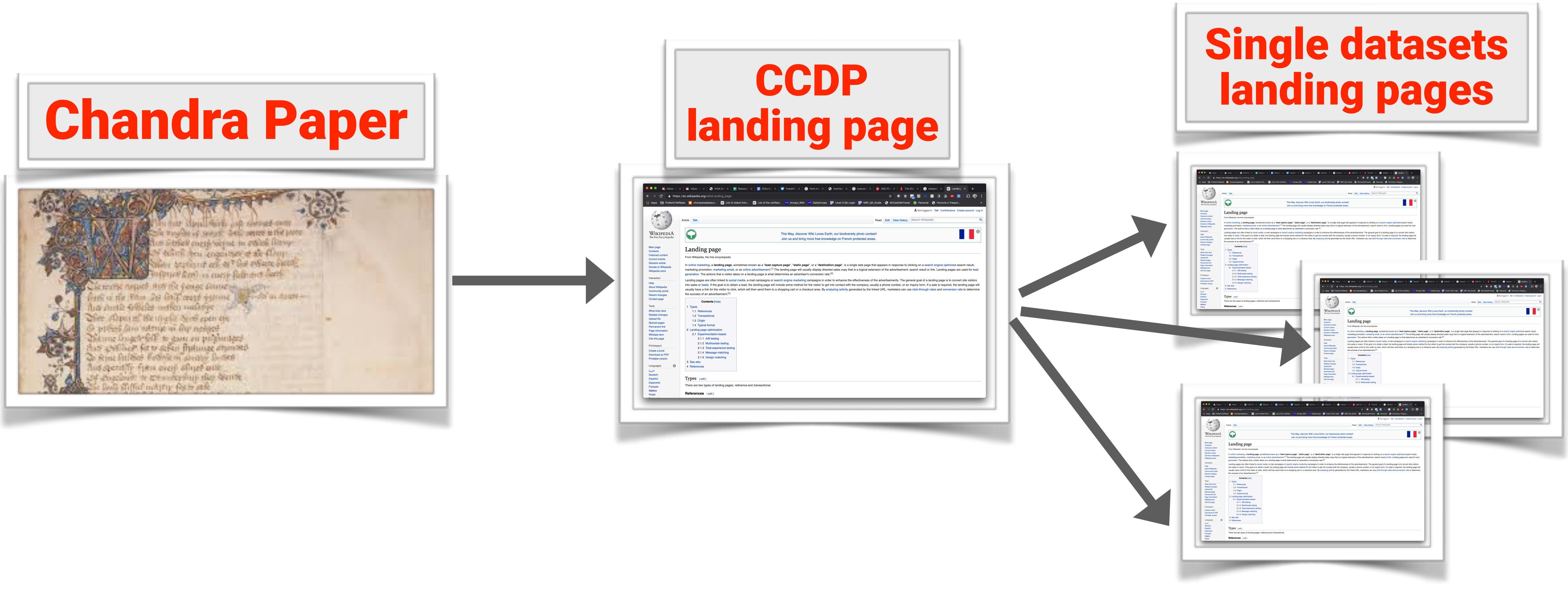
L2 (single ObsIDs)

DOIs for data in the archive

- **Dataset-based DOIs**
  - **single observations** (L2 observations)
  - **aggregates**
    - merged aggregates (catalog-style ***stacks***)
    - “unintentional” spatial aggregates (collections of multiple L2 observations)
  - User-contributed aggregates
    - heterogenous types of data
  - **Chandra Source Catalog DOI**(s)
    - one DOI for each version of the CSC, associated to a landing page that can resolve fragments to achieve the required granularity

Chandra paper collections

- **Chandra paper collections DOIs** (aka Consolidated Chandra Data Page - CCDP)
  - basic bibliographic metadata about each Chandra science paper
  - intermediate landing page linking landing pages to each CDA dataset used in the paper



Chandra Paper

CCDP landing page

Single datasets landing pages

Metadata

- Requirements on metadata assignment/definition
  - accurate enough to allow **unequivocal identification of object**
  - need to express relationships with other associated ***objects***
    - literature objects
    - other related data objects
  - include versioning information
  - including path (landing page) to the data objects

- Requirements on upkeep of metadata
  - "**one-and-done**" metadata
    - observational/data objects metadata that won't change over time…
    - …or change seldom
  - **continuously updated metadata**
    - literature objects that keep using the same data products
    - new types/level of aggregations of basic data products

Example of metadata assignments/updates

Identifier	IdentifierType=DOI *<DOI>*			
titles	title=Chandra X-ray Observatory ObsId *<ObsId>*			

creators	creator	creatorname=CXC-DS
		affiliation=Smithsonian Astrophysical Observatory		

publisher=Chandra X-ray Center/SAO				
publicationYear=*<year data became/will become public>*				
resourceType	resourceTypeGeneral=Dataset Astronomical Data			
subjects	subject=High Energy Astrophysics Data/X-ray Data			

fundingReferences	fundingReference	funderName=NASA
		awardTitle=Chandra X-ray Center
		awardNumber=NAS-8-03060		

contributors	contributor	contributorType=RightsHolder
		contributorName=NASA		

	contributor	contributorType=HostingInstitution
		contributorName=SAO		

	contributor	contributorType=DataManager
		contributorName=ChandraDataArchive		

	contributor	contributorType=RegistrationAgency
		contributorName=Smithsonian Institution		

	contributor	contributorType=Distributor
		contributorName= Chandra Data Archive		

dates	date	dateType=Collected *<observation date in yyyy-mm-dd (START_DATE)>*		
	date	dateType=Created *<V&V date of first distribution in yyyy-mm-dd>*		
	date	dateType=Available *<public release date in yyyy-mm-dd>*		
descriptions	description	descriptionType=Abstract *<proposal title>*		
geoLocations	geoLocation	geoLocationPosition=ICRS		

		geoLocationPoint	pointLongitude=*<RA>*
			pointLatitude=*<Dec>*	

		geoLocationPolygon	polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

			polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

				…
		<polygon describing the first chip or HRC detector>		

		geoLocationPolygon	polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

			polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

			…	
		<polygon describing the second chip ON in ACIS (if any)>		
		…		
sizes	size	*<n>* MB Primary Data Package		
	size	*<m>* MB Secondary Data Package		
	size	*<x>* ks Exposure Time (as given in ChaSeR)		
formats	format	FITS		
version	*<version>*			
rights	Public Data|Proprietary Data			

Metadata Updates

In addition, the following events trigger updates:

Reprocessing

Add:

dates	date	dateType=Updated *<V&V date in yyyy-mm-dd>*		

Update:

geoLocations	geoLocation	geoLocationPosition=ICRS		

		geoLocationPoint	pointLongitude=*<RA>*
			pointLatitude=*<Dec>*	

		geoLocationPolygon	polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

			polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

				…
		<polygon describing the first chip or HRC detector, from fov1.fits file>		

		geoLocationPolygon	polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

			polygonPoint	pointLongitude=*<RA>*
				pointLatitude=*<Dec>*

				…
		<polygon describing the second chip ON in ACIS (if any), from fov1.fits file>		
		…		
sizes	size	*<n>* MB Primary Data Package		
	size	*<m>* MB Secondary Data Package		
	size	*<x>* ks Exposure Time (as in ChaSeR)		
version	*<n>*			

Archiving (becoming public)

rights	Public Data

Aggregation into new aggregates

relatedIdentifiers	relatedIdentifier	relatedIdentifierType=DOI
		relationType=IsSourceOf
		<DOI of data aggregation containing the observation>

	….	

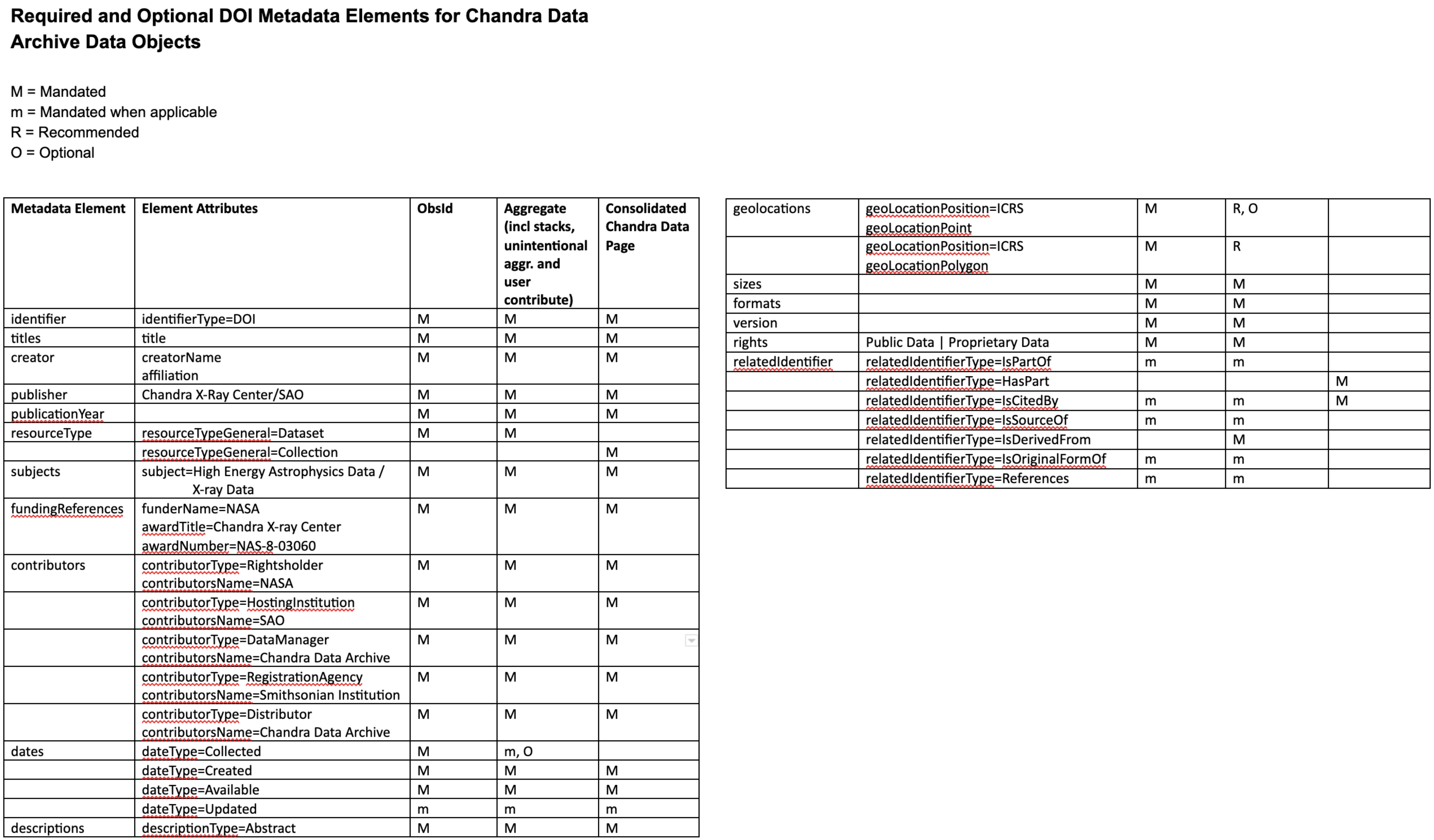
Publications

relatedIdentifiers	relatedIdentifier	relatedIdentifierType=DOI
		relationType=IsPartOf
		<DOI of Consolidated Chandra Data Page>

	relatedIdentifier	relatedIdentifierType=DOI
		relationType=IsCitedBy
		<article DOI>

	relatedIdentifier	relatedIdentifierType=bibcode
		relationType=IsCitedBy
		<article Bibcode>

Example of metadata assignments/updates



Required and Optional DOI Metadata Elements for Chandra Data Archive Data Objects

M = Mandated
m = Mandated when applicable
R = Recommended
O = Optional

Metadata Element	Element Attributes	ObsId	Aggregate (incl stacks, unintentional aggr. and user contribute)	Consolidated Chandra Data Page
identifier	identifierType=DOI	M	M	M
titles	title	M	M	M
creator	creatorName affiliation	M	M	M
publisher	Chandra X-Ray Center/SAO	M	M	M
publicationYear		M	M	M
resourceType	resourceTypeGeneral=Dataset	M	M	
	resourceTypeGeneral=Collection			M
subjects	subject=High Energy Astrophysics Data / X-ray Data	M	M	M

fundingReferences	funderName=NASA
	awardTitle=Chandra X-ray Center
	awardNumber=NAS-8-03060
		M	M	M

contributors	contributorType=Rightsholder
	contributorsName=NASA
		M	M	M

	contributorType=HostingInstitution
	contributorsName=SAO
		M	M	M

	contributorType=DataManager
	contributorsName=Chandra Data Archive
		M	M	M

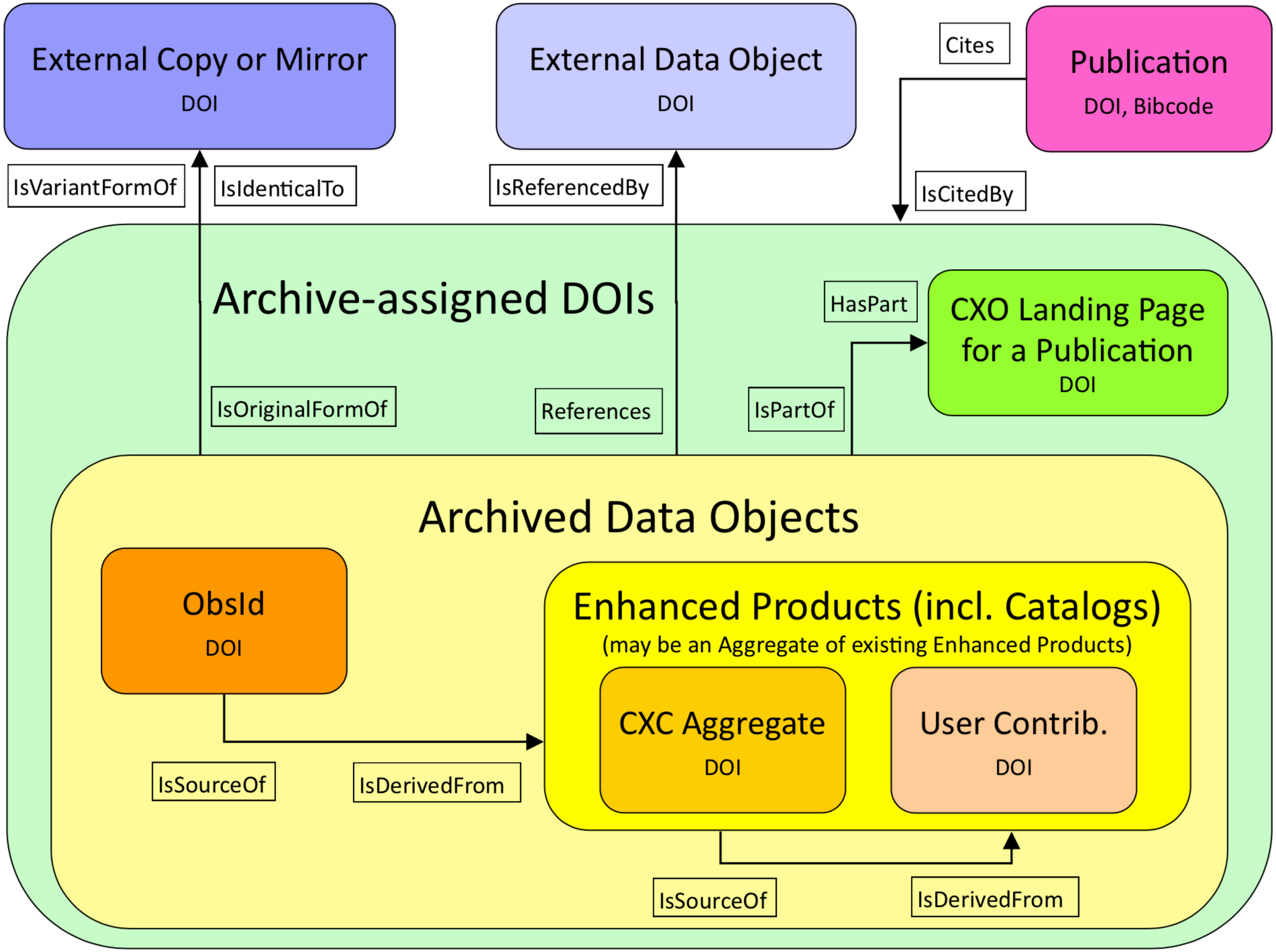
	contributorType=RegistrationAgency
	contributorsName=Smithsonian Institution
		M	M	M

	contributorType=Distributor
	contributorsName=Chandra Data Archive
		M	M	M

dates	dateType=Collected	M	m, O	
	dateType=Created	M	M	M
	dateType=Available	M	M	M
	dateType=Updated	m	m	m
descriptions	descriptionType=Abstract	M	M	M
geolocations	geoLocationPosition=ICRS geoLocationPoint	M	R, O	
	geoLocationPosition=ICRS geoLocationPolygon	M	R	
sizes		M	M	
formats		M	M	
version		M	M	
rights	Public Data | Proprietary Data	M	M	
relatedIdentifier	relatedIdentifierType=IsPartOf	m	m	
	relatedIdentifierType=HasPart			M
	relatedIdentifierType=IsCitedBy	m	m	M
	relatedIdentifierType=IsSourceOf	m	m	
	relatedIdentifierType=IsDerivedFrom		M	
	relatedIdentifierType=IsOriginalFormOf	m	m	
	relatedIdentifierType=References	m	m	

Relational identifiers

CDA relationTypes for relatedIdentifiers



Practical considerations

- ➡ DataCite metadata schema 4.1 provides flexibility to define (a very basic set of) properties of Chandra observations
- ➡ SI is a DataCite member, SAO can mint DOIs with the prefix 10.0344
  - ➡ backfilling of the archive: ~40,000 DOIs
  - ➡ average number of new DOIs: ~3,000/year
  - ➡ creating mechanism to generate landing pages for all classes of data products
- ➡ DOIs will replace the *ivo* identifiers currently used
  - ➡ the *ivo* identifiers populate the DS_IDENT keyword in FITS headers
    - ➡ DS_IDENT= 'ADS/Sa.CXO#obs/22056' / dataset identifier -> DS_IDENT= '10.0344/SAO.CXO.obs.22056'
  - ➡ CIAO tool *list_datasetid* reads, creates and lists PIDs for Chandra observations
  - ➡ dependencies on VO protocols!