Cheap and FAIR:

Building a Serverless Research Data Repository

Identifiers & Landing Pages

Andrea Zonca & Rick Wagner
zonca@sdsc.edu & rick@sdsc.edu
SDSC

Gateways 2024, September 30, 2024, Online

Identifiers…*hoo boy*

- Identifiers can be a contentious topic
- It’s all about namespaces and resolvers
- Identifiers don’t need to be semantically meaningful
- Datasets can have multiple identifiers
    - E.g., You can start with a local identifier and then add a DOI
- URLs are challenging
    - Using a URL for an identifiers means the identifier “owns” a URL

Identifiers: Some Reasonable Terms

identifier: a chain of characters used to refer to an entity

locally unique identifier: an identifier which is unique to a given context but which may clash if moved out of this specific context

globally unique identifier: an identifier which is produced such that the probability of the exact same string is extremely low (but not null), hence global—AKA a universally unique identifier (UUID)

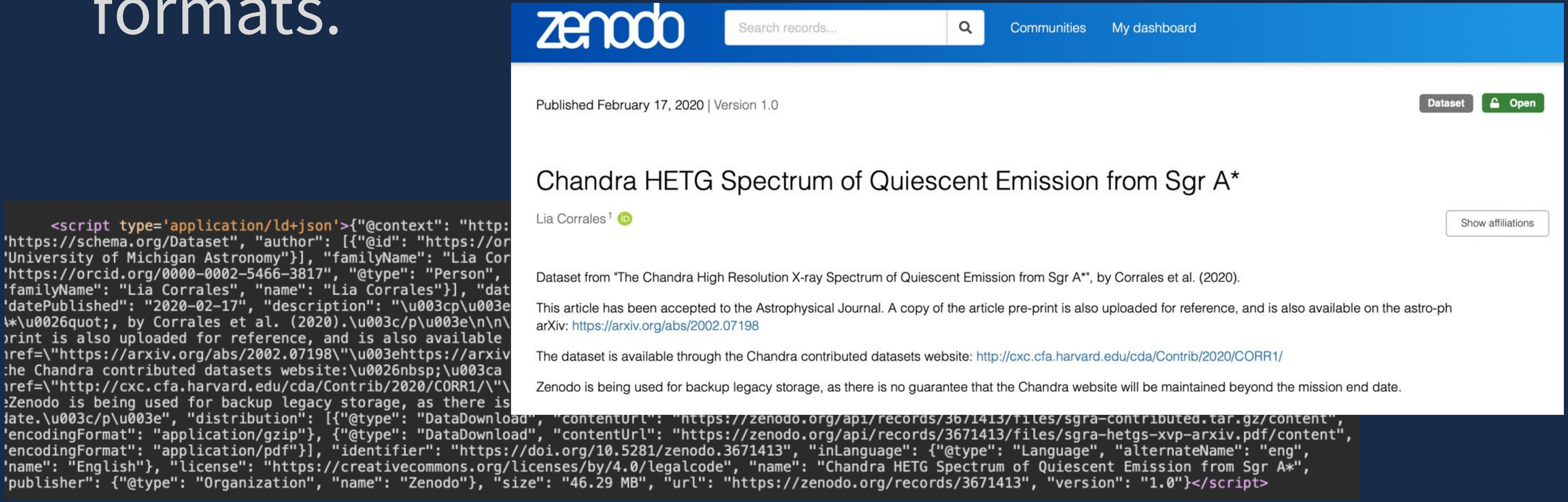
persistent identifier: an identifier which provides a long-lasting reference to a digital resource—needs a resolution service (e.g., DOI)

identifier resolution service: software infrastructure that can return a URL (i.e., a *landing page*) for an identifier (e.g., DataCite)

https://nih-cfde.github.io/the-fair-cookbook/content/recipes/Identification/identifiers.html

Landing Pages

A landing page is the web page that describes a dataset. It provides the metadata about the dataset in human & machine-readable formats.



Published February 17, 2020 | Version 1.0

Chandra HETG Spectrum of Quiescent Emission from Sgr A*

Lia Corrales[1]

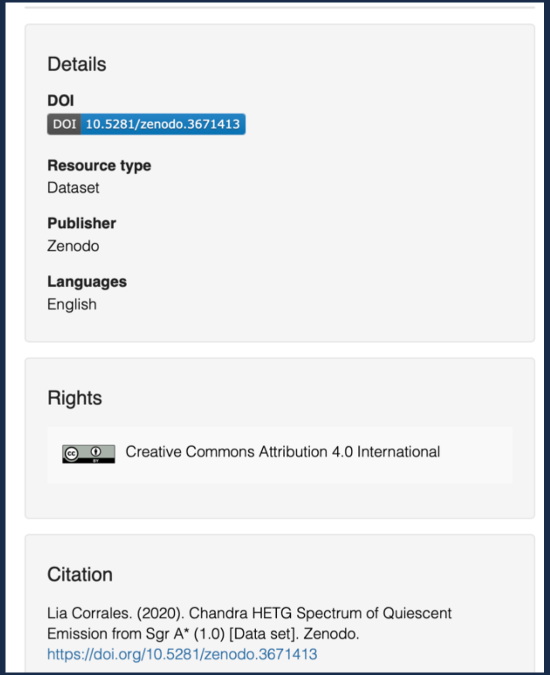
Dataset from "The Chandra High Resolution X-ray Spectrum of Quiescent Emission from Sgr A*", by Corrales et al. (2020).

This article has been accepted to the Astrophysical Journal. A copy of the article pre-print is also uploaded for reference, and is also available on the astro-ph arXiv: https://arxiv.org/abs/2002.07198

The dataset is available through the Chandra contributed datasets website: http://cxc.cfa.harvard.edu/cda/Contrib/2020/CORR1/

Zenodo is being used for backup legacy storage, as there is no guarantee that the Chandra website will be maintained beyond the mission end date.

```
     <script type='application/ld+json'>{"@context": "http:
'https://schema.org/Dataset", "author": [{"@id": "https://or
'University of Michigan Astronomy"}], "familyName": "Lia Cor
'https://orcid.org/0000-0002-5466-3817", "@type": "Person",
'familyName": "Lia Corrales", "name": "Lia Corrales"}], "dat
'datePublished": "2020-02-17", "description": "\u003cp\u003e
A*\u0026quot;, by Corrales et al. (2020).\u003c/p\u003e\n\n\
print is also uploaded for reference, and is also available
href=\"https://arxiv.org/abs/2002.07198\"\u003ehttps://arxiv
the Chandra contributed datasets website:\u0026nbsp;\u003ca
href=\"http://cxc.cfa.harvard.edu/cda/Contrib/2020/CORR1/\"\
eZenodo is being used for backup legacy storage, as there is
date.\u003c/p\u003e", "distribution": [{"@type": "DataDownload", "contentUrl": "https://zenodo.org/api/records/3671413/files/sgra-contributed.tar.gz/content",
'encodingFormat": "application/gzip"}, {"@type": "DataDownload", "contentUrl": "https://zenodo.org/api/records/3671413/files/sgra-hetgs-xvp-arxiv.pdf/content",
'encodingFormat": "application/pdf"}], "identifier": "https://doi.org/10.5281/zenodo.3671413", "inLanguage": {"@type": "Language", "alternateName": "eng",
'name": "English"}, "license": "https://creativecommons.org/licenses/by/4.0/legalcode", "name": "Chandra HETG Spectrum of Quiescent Emission from Sgr A*",
'publisher": {"@type": "Organization", "name": "Zenodo"}, "size": "46.29 MB", "url": "https://zenodo.org/records/3671413", "version": "1.0"}</script>
```



Details

DOI

DOI 10.5281/zenodo.3671413

Resource type

Dataset

Publisher

Zenodo

Languages

English

Rights

Creative Commons Attribution 4.0 International

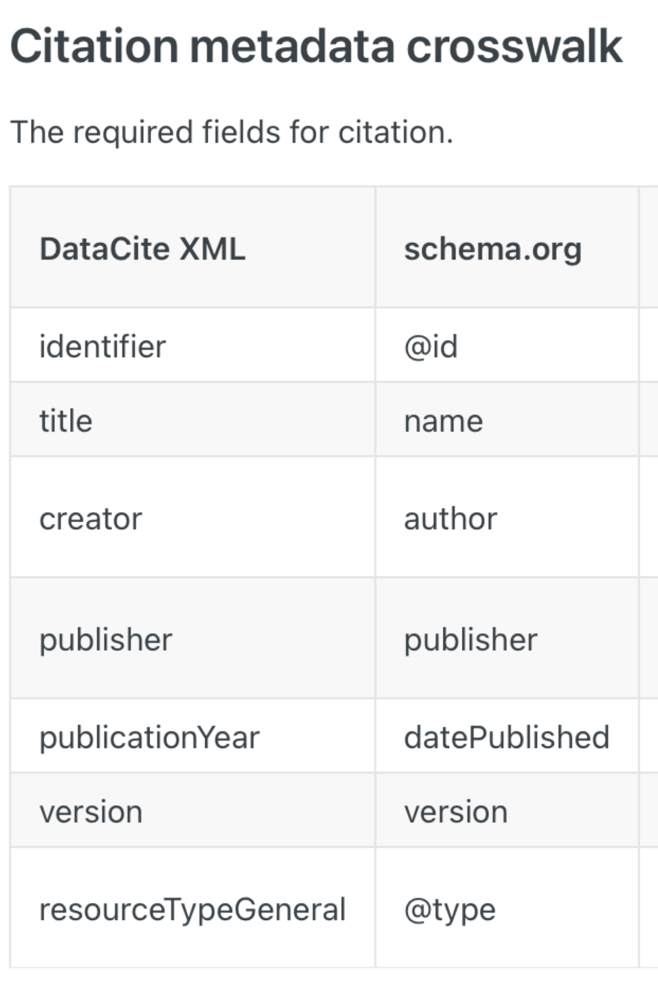
Citation

Lia Corrales. (2020). Chandra HETG Spectrum of Quiescent Emission from Sgr A* (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3671413

Two Metadata Standards to Know

DataCite	Schema.org
Used for DOIs	Used for Google Dataset Search

- Use clearest terms on your landing pages (human-readable)
- Use Schema.org to generate JSON-LD embedded in your landing pages (machine-readable)



Citation metadata crosswalk

The required fields for citation.

DataCite XML	schema.org
identifier	@id
title	name
creator	author
publisher	publisher
publicationYear	datePublished
version	version
resourceTypeGeneral	@type

Recommended Practices

- Assign every dataset an identifier unique to your repository
- Use this identifier in your metadata
- Ensure that every dataset landing page
  - has the minimum DataCite and Schema.org fields
  - in both JSON-LD and human-readable formats
- Landing pages are public even when the data is not
- Use a Creative Commons or OSI license if possible
- If you assign a DOI to your dataset, be diligent about maintaining the landing page
- Landing pages need to stay even if the data is gone