FAIR SSH Data Citation

A Practical Guide

Nicolas Larrousse, Huma-Num CNRS
Edward J. Gray, Huma-Num CNRS
Cesare Concordia, ISTI-CNR
Daan Broeder , CLARIN ERIC

SSHOC T3.4 “Making Data Findeable by being Citable”

3 December 2021

Online

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

HOUSEKEEPING NOTES

- **The webinar is being recorded.** All participants will receive a link to the recording shortly after the event.
  - Please keep your camera and microphone off, if you do not wish to appear on the recording.
- **Slides are available:** See the chat box for the link.
- **Questions:** Write them in the chatbox or ask them during the Q&A session.
- **Post-event feedback:** https://forms.gle/9VMN99YizicUG6RR8

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

SPEAKERS

Nicolas Larrousse
Huma-Num/CNRS

Edward J. Gray
Huma-Num/CNRS

Cesare Concordia
ISTI-CNR

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Project: SSHOC
social sciences & humanities open cloud

Horizon 2020
European Union Funding
for Research & Innovation

Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 47
(20 beneficiaries + 27 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Duration: 40 months
(January 2019 – 30 April 2022)

Project budget:
€ 14,455,594.08

Project website:
www.SSHOpenCloud.eu

Objectives:

- creating the social sciences and humanities **(SSH)** part of European Open Science Cloud **(EOSC)**
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

SSHOC Partners

ESFRI Landmarks + projects

cessda · DARIAH-EU · CLARIN · European Social Survey · SHARE ERIC Survey of Health, Ageing and Retirement in Europe · E-RIHS European Research Infrastructure for Heritage Science *

Stakeholder engagement & dissemination

LIBER · Trust-IT Services Communicating ICT to markets

Research Communities

University of York · SciencesPo · Koninklijke Nederlandse Akademie van Wetenschappen · cnrs · Consiglio Nazionale delle Ricerche · University of Nottingham UK | China | Malaysia · Tilburg University · The National Gallery · UCL

Technology providers

FORTH Institute of Computer Science · Semantic Web Company · CentERdata

*E-RIHS is not a legal partner in the SSHOC project but we connect to the E-RIHS community through the Institutum Archaeologicum Germanicum.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Examples of SSHOC topics and activities

- Strengthening and certification of SSH data repositories
- The use of language technology for the Social Sciences (Machine Translation) for the translation of surveys
- Generalisation of services for use by all the SSH
- A SSH Open Marketplace for discovery of SSH services and data by researchers
- Remote access to sensitive data
- Alignment of SSH data-management practices in line with Open Science & FAIR principles
- Providing trainings and training materials

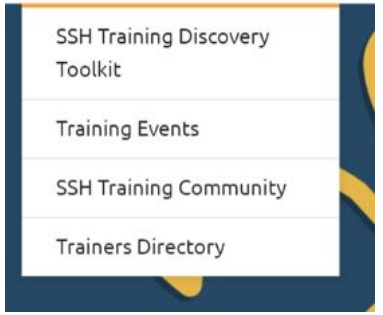
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Some SSHOC offerings to test!

https://www.sshopencloud.eu/ssh-open-marketplace

https://www.sshopencloud.eu/training



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

FUTURE SSHOC TRAINING EVENTS

- 24 and 25 Jan 2022: SSHOC Workshop: Copyright Issues in Secondary Data Use
  - https://tinyurl.com/vy6xwydx
- 14 and 15 Feb 2022: SSHOC Workshop: Data Management Planning and Overcoming Challenges in Social Sciences Data Sharing
  - *https://tinyurl.com/8mwbkjv2*
- Follow the SSHOC channels to be informed about registration details.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

PART 1: Data Citation in SHS

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Data Citation in SSH?

- Reproducibility and transparency of the research process
- Give credit to the creator and the funder of the data
- Provide confidence in the data and the context of their production
- Give visibility
- Prove the usefulness of infrastructures
- Reuse data for different research purposes in other contexts

Etc.

In the general context of development of links between Data & Publications / Data papers / Data journals? …

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Current Situation in SSH

- Very diverse and no specific common approach to data citation
- The notion of “publishing data” is relatively new
- Social Sciences have a long tradition of data citation

-> Requirements exist (DASISH Project, ICPSR, CESSDA, SHARE, W3C’s Web Annotation Data Model, RDA Data Citation of Evolving Data etc.)

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Data Citation in SSH: Some practical examples

LDOR, & Thésaurus Occitan. (2015). Atlas Linguistique et ethnographique du Languedoc Occidental. [Data set]. Cocoon. https://doi.org/10.34847/cocoon.d7c25365-6234-33ef-b4fb-e01029a23c47

```
@misc{https://doi.org/10.34847/cocoon.d7c25365-6234-33ef-b4fb-e01029a23c47,
  doi = {10.34847/COCOON.D7C25365-6234-33EF-B4FB-E01029A23C47},
  url =
{https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-d7c25365-6234-33ef-b4fb-e01029a23c47},
  author = {{LDOR} and {Thésaurus Occitan}},
  keywords = {oci},
  language = {oc},
  title = {Atlas Linguistique et ethnographique du Languedoc Occidental},
  publisher = {Bases, corpus, langage; Équipe de Recherche en Syntaxe et Sémantique},
  year = {2015}
}
```

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Our journey to citation landscape: how we got there?

Recommendations for FAIR Data Citation in the Social Sciences and Humanities

link

Consultation in events & expert review

Inventory of current practices about Data Citation within SSH

Deliverable 3.2

Review of SSH Data Repositories

Deliverable 3.5

Data Citation Prototype

link

Automatically extracts citation metadata from PIDs (persistent identifiers) and other sources

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

PART 2: Data Citation Recommendations and Survey of Repositories

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Recommendations

- Based on Force11 8 Data Citation Principles[1]
- Further developed thanks to peer review and at a Round Table of Experts this May

Explanations of the general principles of data citation in SSH. First explains the general, societal challenge towards data citation, then the recommendation to solve this, and the expected outcome.

Societal/Technical Challenge (adapted from FORCE11 principles)	Recommendation	Expected Outcomes
Persistence: Research data should persist beyond the research project itself. Until recently, SSH data was not considered as a crucial product of the research process, more as a tool used to conduct research.	➢ Create and maintain sustainable infrastructures for SSH in order to achieve persistence ➢ Use trusted data repositories with a clear roadmap and good practices that comply with standards (e.g.,long-term preservation or accessibility, etc.) ➢ Train researchers to build a DMP at the very beginning of the project with the support of data stewards and periodically update it during the project lifecycle (i.e. living document) ➢ Support researchers in the execution of their data management strategy ➢ Only research data that is citable is to be preserved	➢ Enhancing discoverability, identification, accreditation and potential reuse of data ➢ Improving the preservation of research data to help justify and off-set the costs of producing it

1. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Use cases for FAIR Data Citation in SSH

I am a Researcher and/or an Engineer working for a project. *As it is becoming more commonplace, and at times required (e.g.,by funders and institutions), to draft Data Management Plans (DMPs), I put my processed research data in trusted Open Science repositories to open my data for potential reuse, allow for citation of my work, and enhance trust in my research. I also want to have an idea of who uses my datasets and how.*
I am a Research software engineer working for a project and/or research infrastructure. *Since I am involved in continuous data collection and handling by means of the software I am maintaining, I am also interested in the tools other researchers use for handling the dataset.*
I am a Manager of a data repository. *I want to understand the use and citation of the research data hosted by my repository, so that I can show qualitative and quantitative figures (e.g., to my funders).*
I am a Data Librarian, a Data Steward or an Open Science Officer. *I support the work of researchers and provide guidance for best practices for research data citation and reuse in their research projects.*
I am a Research Funder*. I want to have a clear view of the "degree of compliance" regarding the DMP submitted by the research project. I want to have a "citation index" of datasets financed to demonstrate the impact of our investment or identify understudied or underfunded research subjects.*
I am a Researcher who is conceiving a research project. *I wish to see what has already been done as I investigate the feasibility of a future project, as a sort of "data bibliography" to understand what research data already relates to the subject or to reuse existing data.*
I am a member of the public who wishes to reuse data. *Either in my work as a journalist, data scientist, or simply an interested citizen, I can find and reuse datasets via proper data citation.*

Recommendations

Audience for the Recommendation include all stakeholders in Data Citation in SSH, from researchers to engineers, and funders to research infrastructures.

Check them out yourself!

Nicolas Larrousse, & Edward J. Gray. (2021). Recommendations for FAIR Data Citation in the Social Sciences and Humanities. Zenodo. https://doi.org/10.5281/zenodo.5361718

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

From recommendations to evaluation of repositories

Once the FAIR SSH Data Citation Recommendations were written and published, we then selected the most pertinent criteria for data repositories and proceeded to evaluate SSH data repositories identified by SSHOC colleagues.

- 85 repositories from SSH landscape (CESSDA, CLARIN, DARIAH and other environments)
- Checks for:
  - Presence of PID, and PID type
  - Presence of Landing Page
  - Presence of Structured Metadata encoded in webpage
  - Presence of “Cite As” Feature
  - Use of Standardized Vocabularies (such as ORCID)
  - Use of Versioning
  - Presence of Links to Related Publications

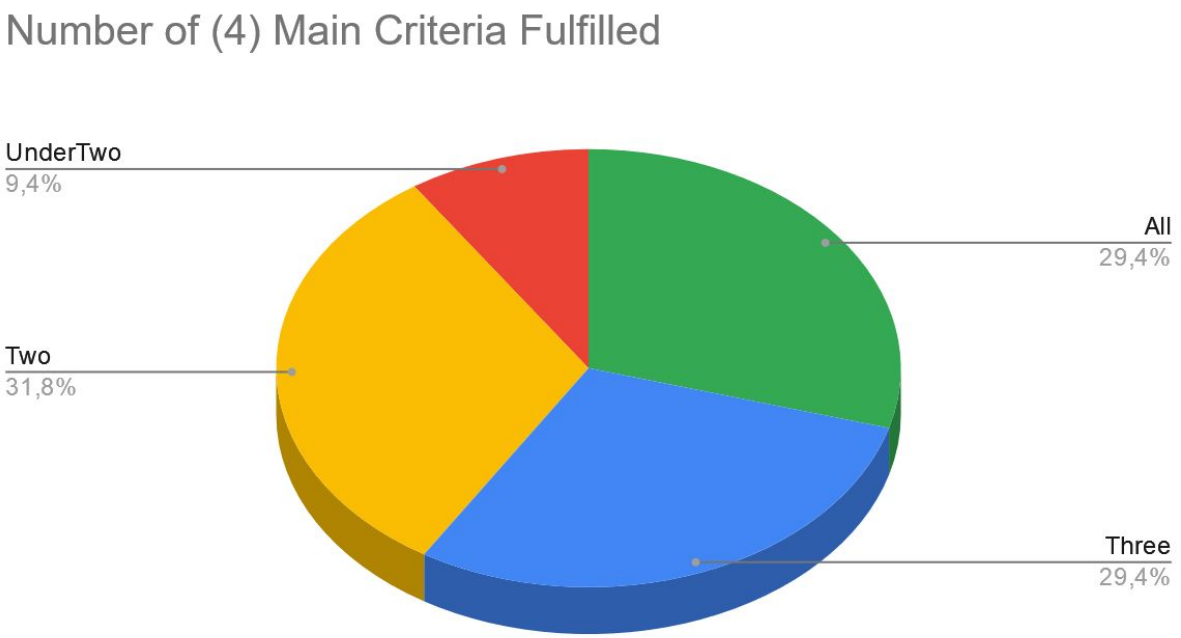
Main Criteria

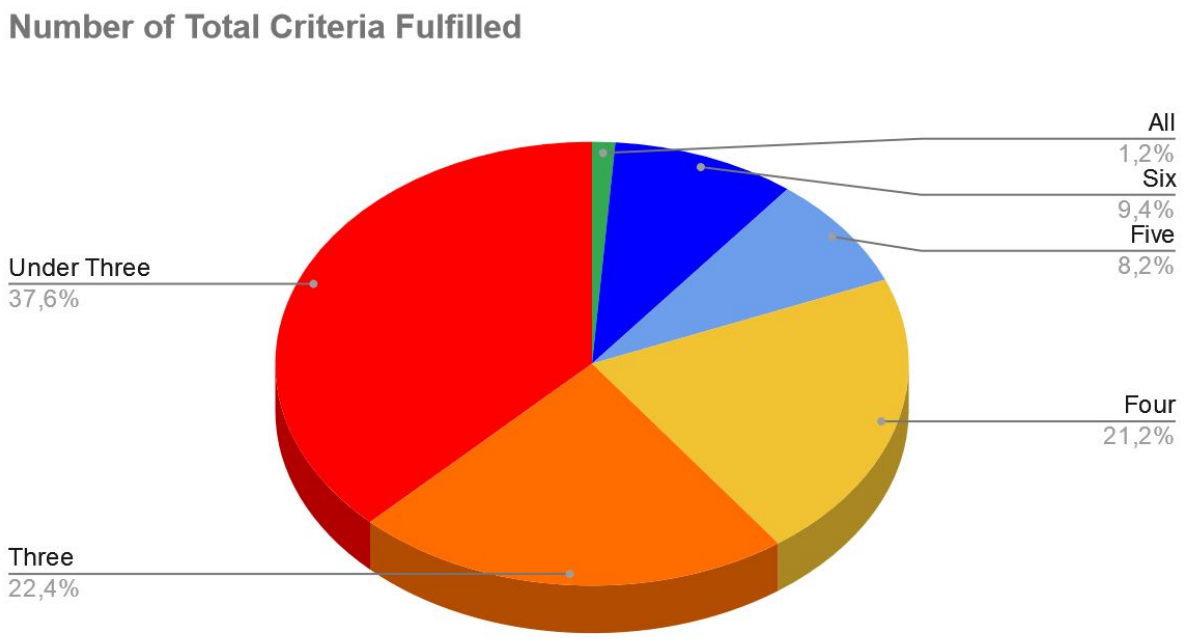
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

From recommendations to evaluation of repositories: Results

Results published in Nicolas Larrousse, Edward Gray, Daan Broeder, Cesare Concordia, Jan Brase, & Athina Papadopoulou. (2021). D3.5 Report on citation enabled SSH catalogues and SSH citation exploitation (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5603306





From recommendations to evaluation of repositories: Results

Overall, results are encouraging, but there is still much work to do to create a fully operational data citation environment:

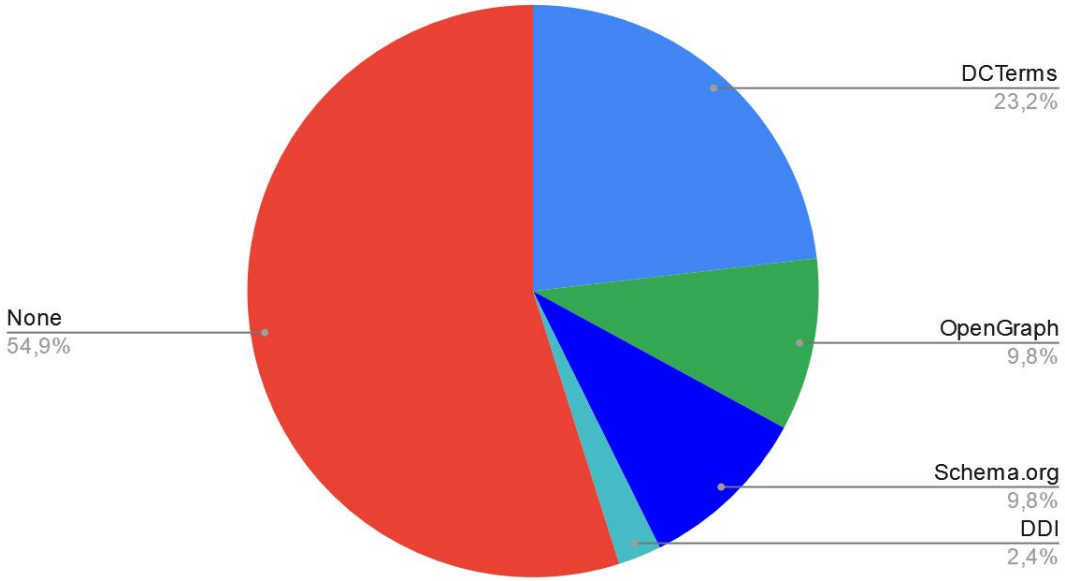
- 87% of repositories had some form of persistent identifier (PID) associated with data
- CiteAs functionality is present in just under half of repositories, yet they are of varying quality
  - A string is not equal to embedded metadata files that can adapt to a chosen standard
- Need better, structured information in landing pages to permit machine actionability
- Versioning only accounted for in 23% of repositories, and links to related publications only in 20%

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

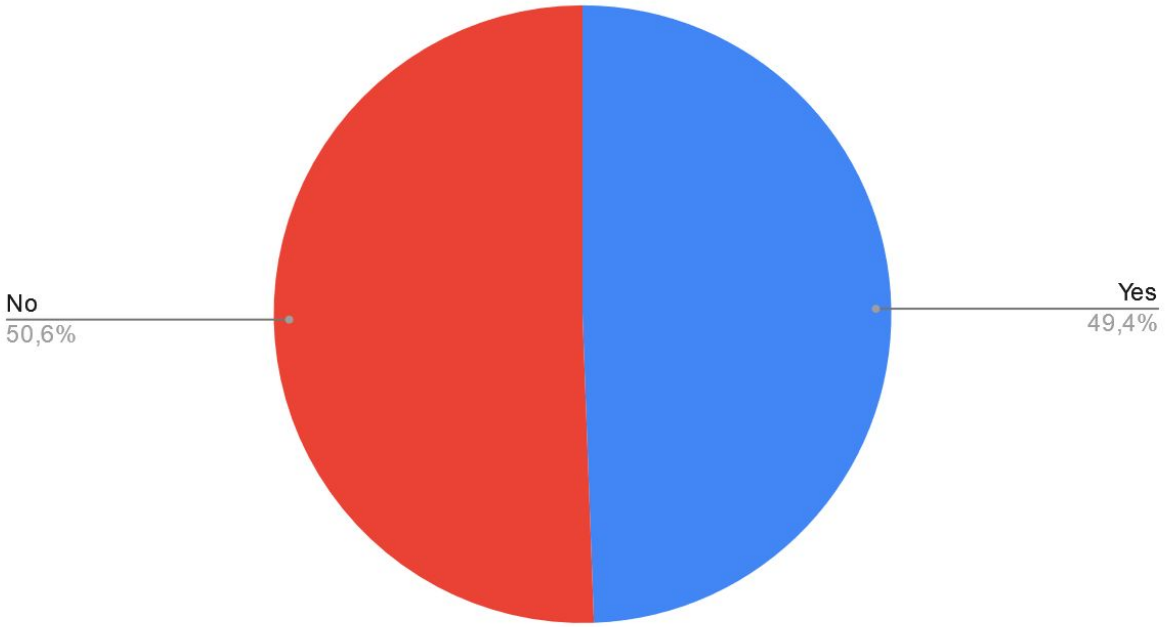
SSHOC
social sciences & humanities open cloud

Main Criteria

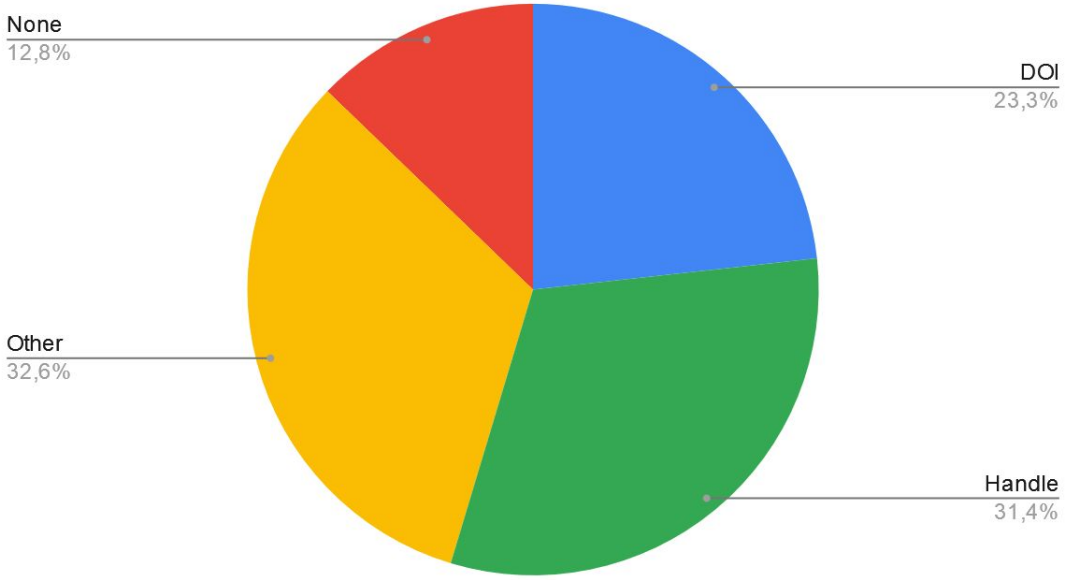
Presence of (embedded) Structured Metadata



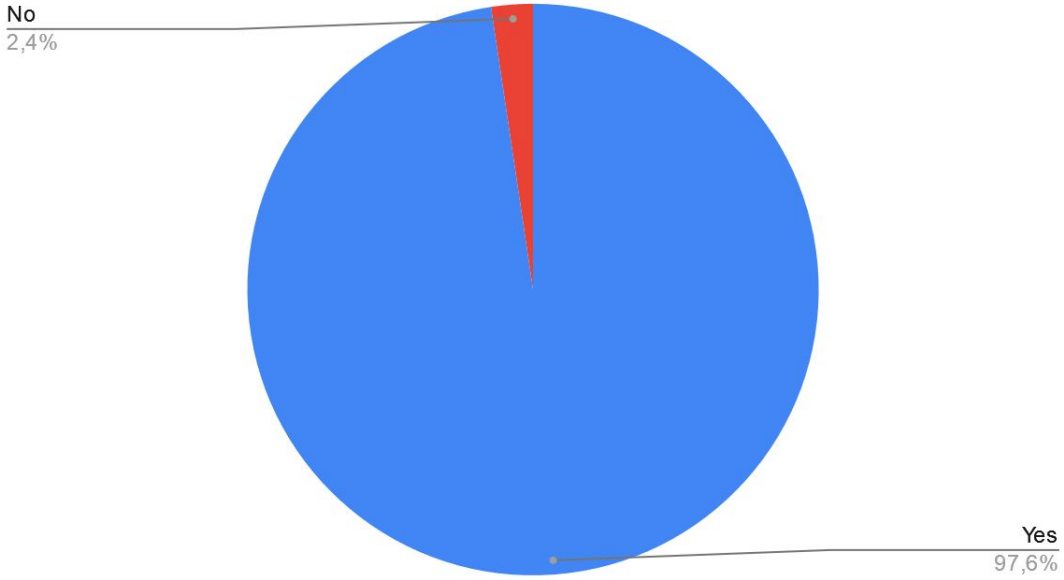
Ready to use "Cite As"



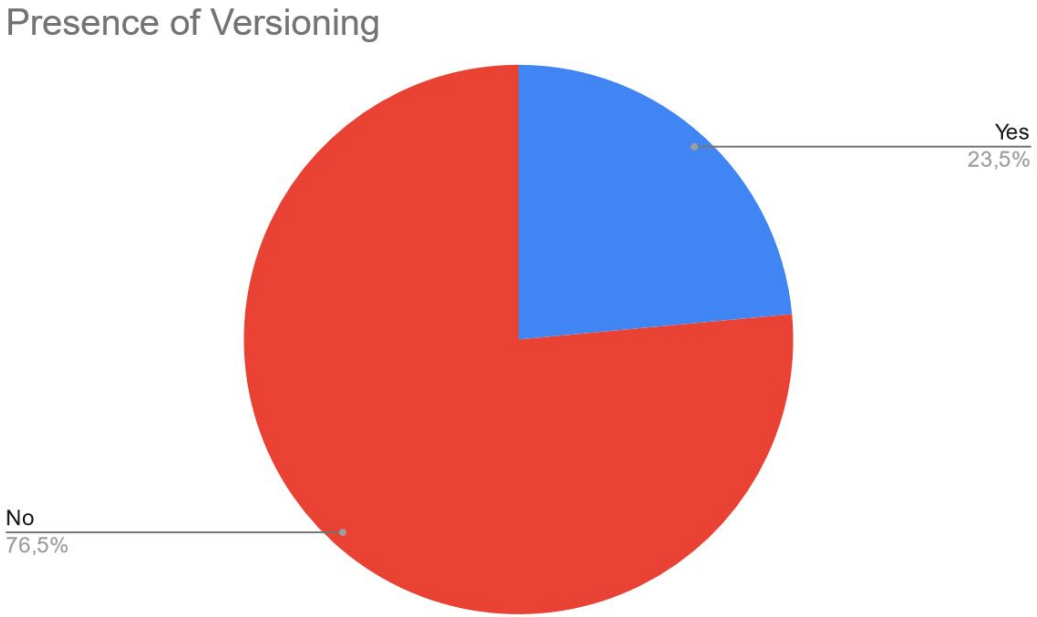
Use of Persistent Identifiers

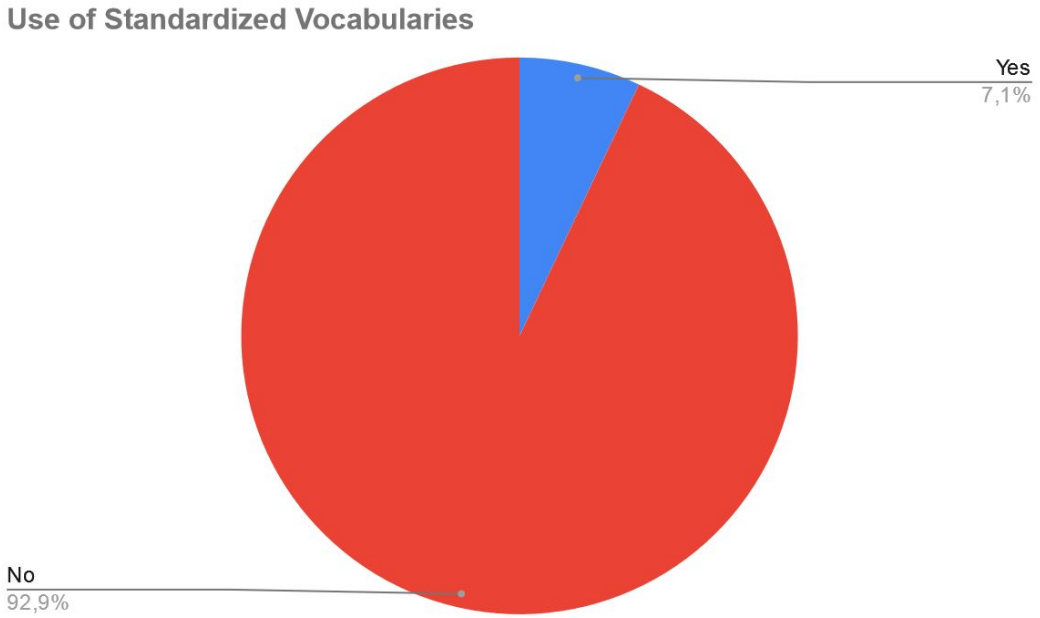


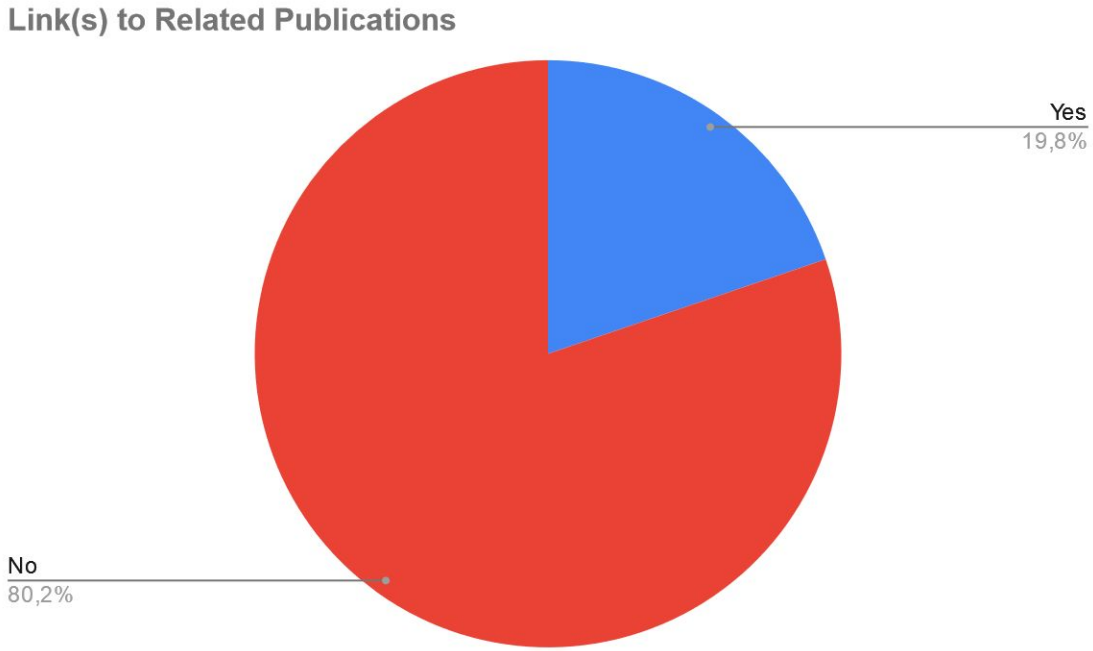
Presence of a Landing Page



Secondary Criteria







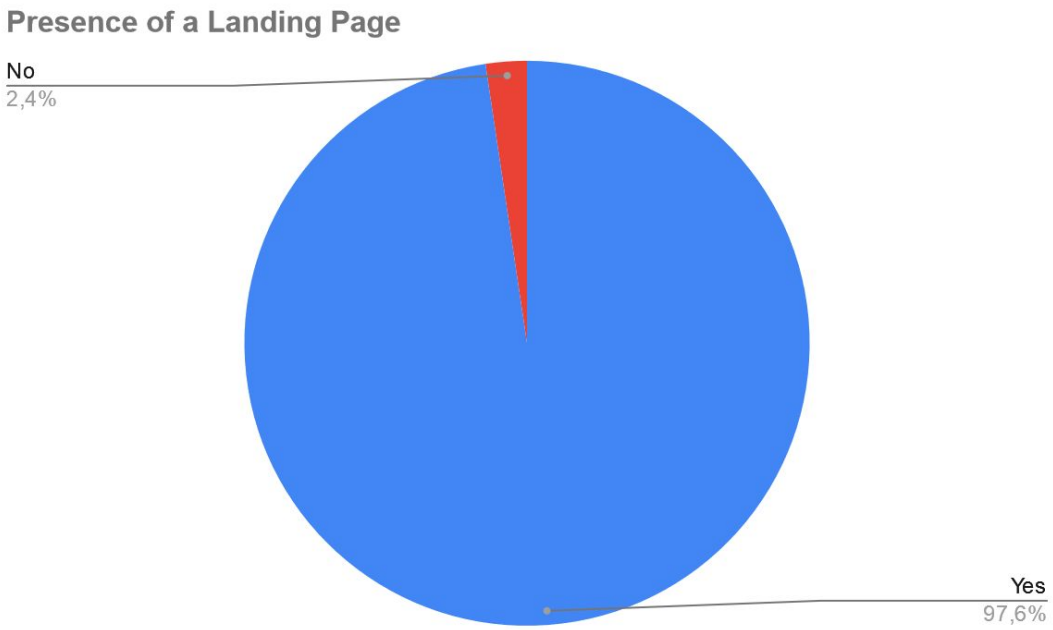
) under Grant Agreement No. 823782

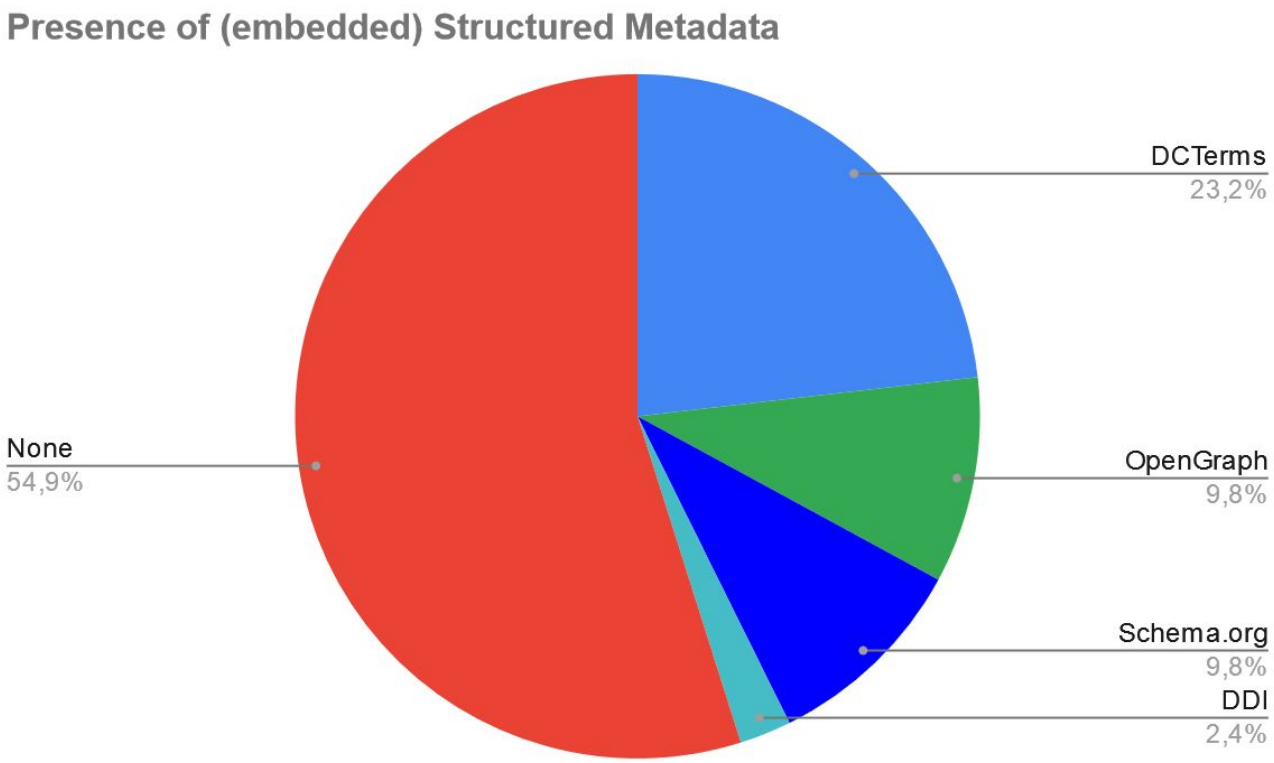
SSHOC
social sciences & humanities open cloud

Landing Pages - keys to machine actionability

If nearly all of the surveyed repositories have a Landing Page, only 45% have structured metadata on those landing pages.

- Structured metadata allow for automatic harvesting of the associated metadata - which can create an even more powerful citation environment.





How can you use these results in practice to build good citations

Before the project

- Read the Recommendations for FAIR Data Citation in SSH and develop your Data Management Plan Accordingly

During the project

- Ensure metadata quality during the life of your project

At the end of the project

- Place your data in a trusted data repository to enable it's use and reuse

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

PART 3 Data Citation Prototype

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Data Citation in practice:
The prototype developed in task 3.4

Getting informations from PIDs (DOIs vs handles/other)
Getting information from landing pages

Getting information from other sources RE3Data / APIs

Gathering information in a standardized way

Citation viewer and Dissemination through an API

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Data Citation in practice: the prototype developed in task 3.4

Prototype key functionalities:

- Explore datasets metadata
  - getting metadata from landing pages/API using PIDs or URLs
  - getting information from other sources
- Provide facilities for curation and semantic annotation of citations.
- Visualize and exploit citations metadata
- Disseminate metadata

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

PIDs used in Repositories

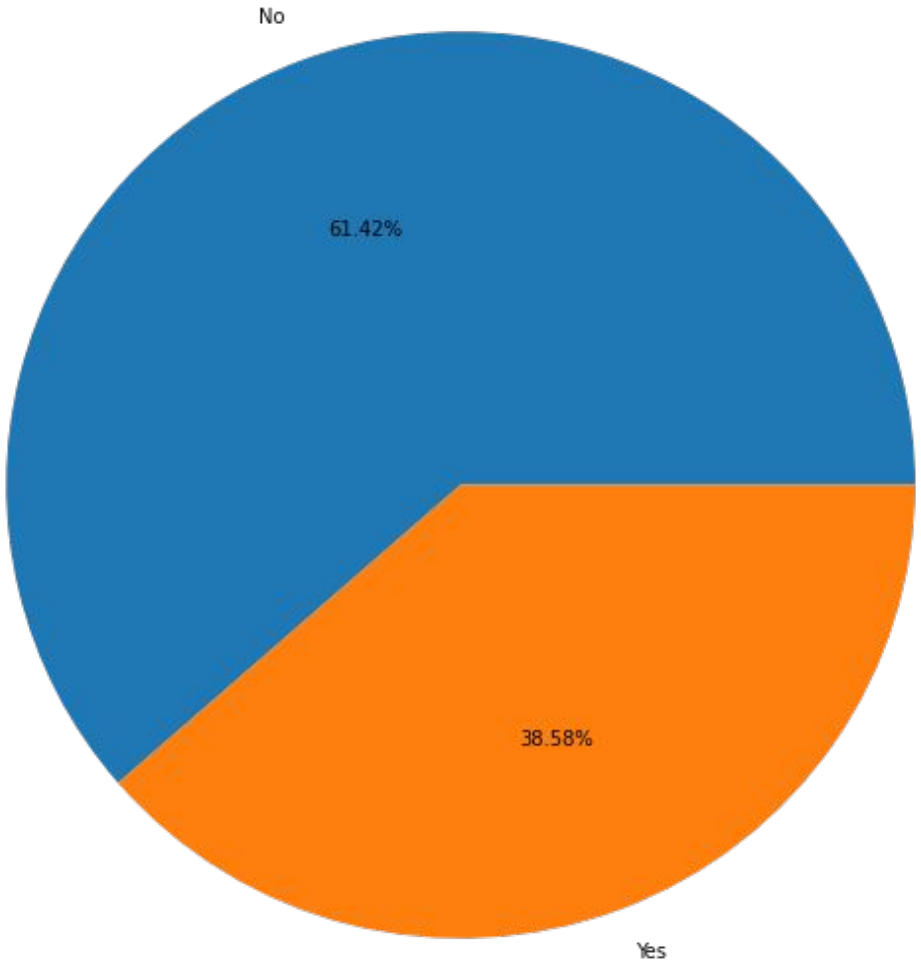
PID	Repositories
None	50
HDL	35
DOI	24
URN	4
URI	2
PURL	1
Permalink	1
Other (local)	1

Source: SSHOC survey and R3Data

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Repositories providing data and metadata via API



Source: R3Data

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Metadata Standards used in repositories

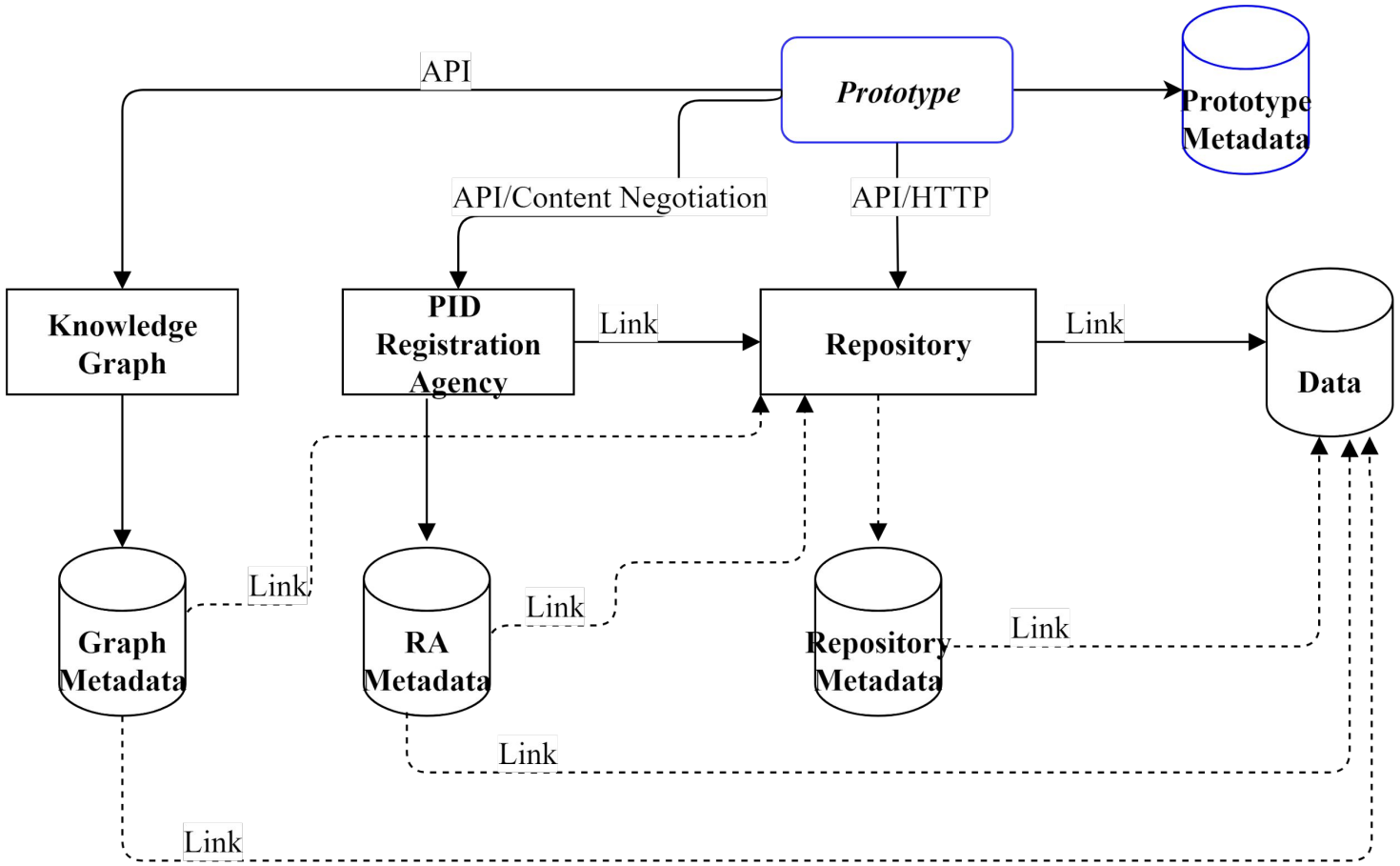
Metadata Standard	Repositories
None	87
Dublin Core	29
DDI - Data Documentation Initiative	15
DataCite Metadata Scheme	5
Repository-developed Metadata Schema	4

Source: R3Data

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Citation Service Prototype: getting metadata from citations



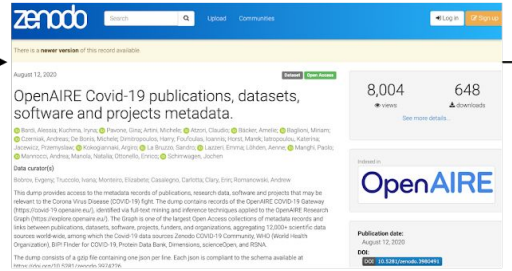
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Getting Metadata: example

10.5281/zenodo.3980491

Citation Prototype



Embedded Metadata

Repository Metadata
license
creator
og:site_name
id
description
type
datePublished
name
url
citation_doi
identifier
citation_publication_date
citation_title
og:title
citation_author
citation_abstract_html_url
distribution
og:description
contributor
og:url

RA Metadata

copyright
author
publisher
id
abstract
type
issued
title
version
URL
DOI

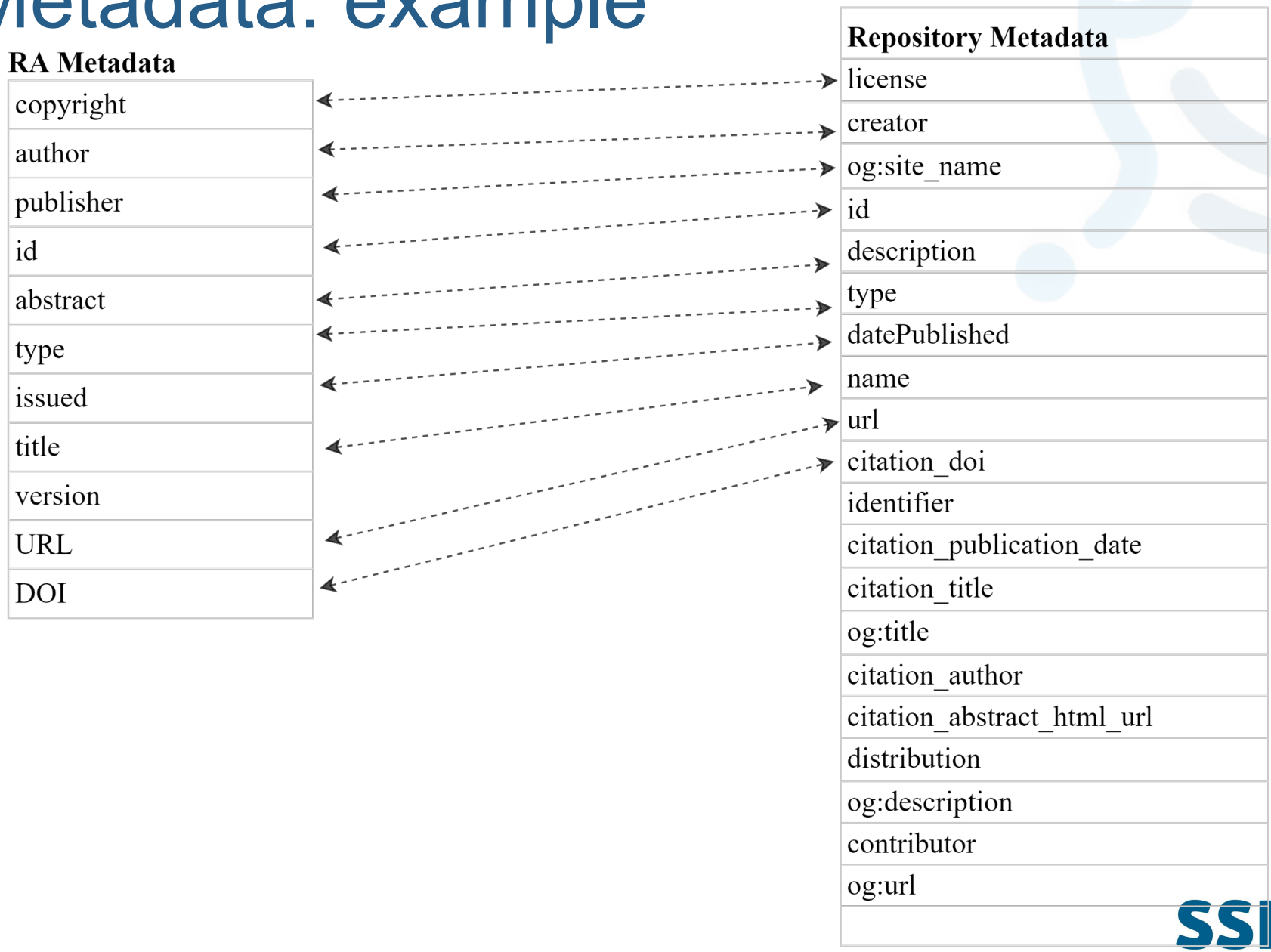
DOI RA

Repository

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Getting Metadata: example



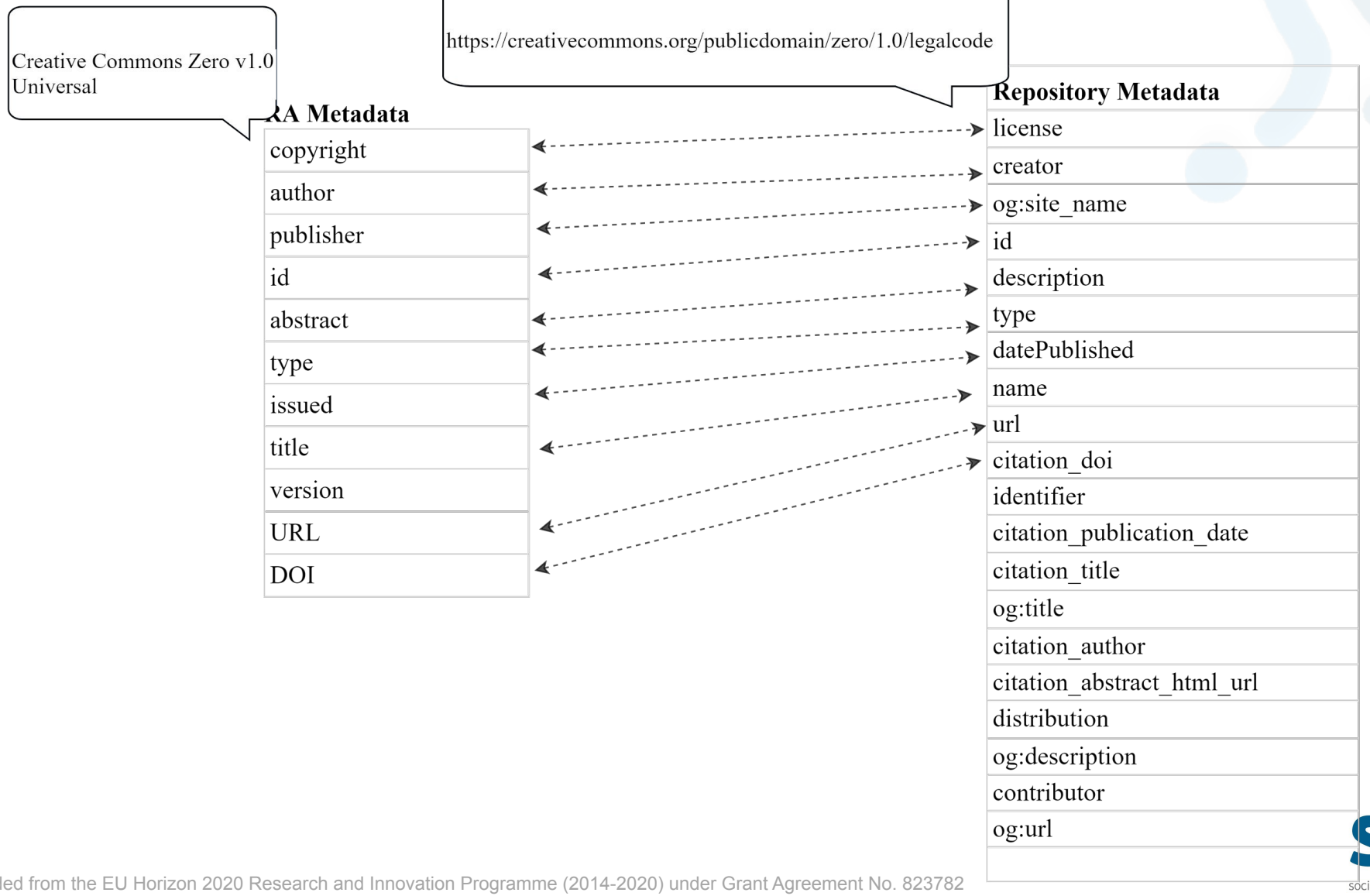
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Getting Metadata: example

Creative Commons Zero v1.0 Universal

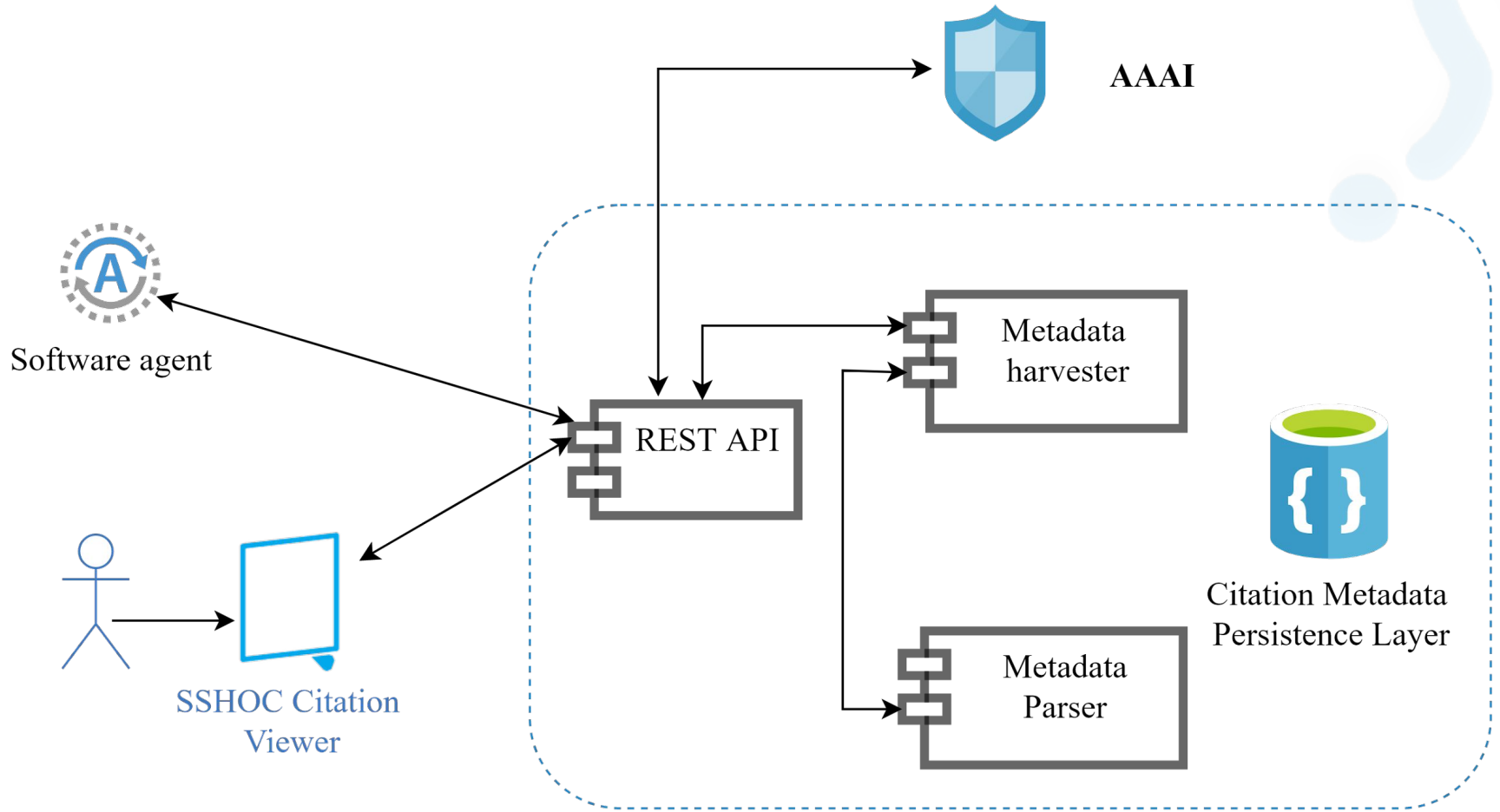
https://creativecommons.org/publicdomain/zero/1.0/legalcode



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

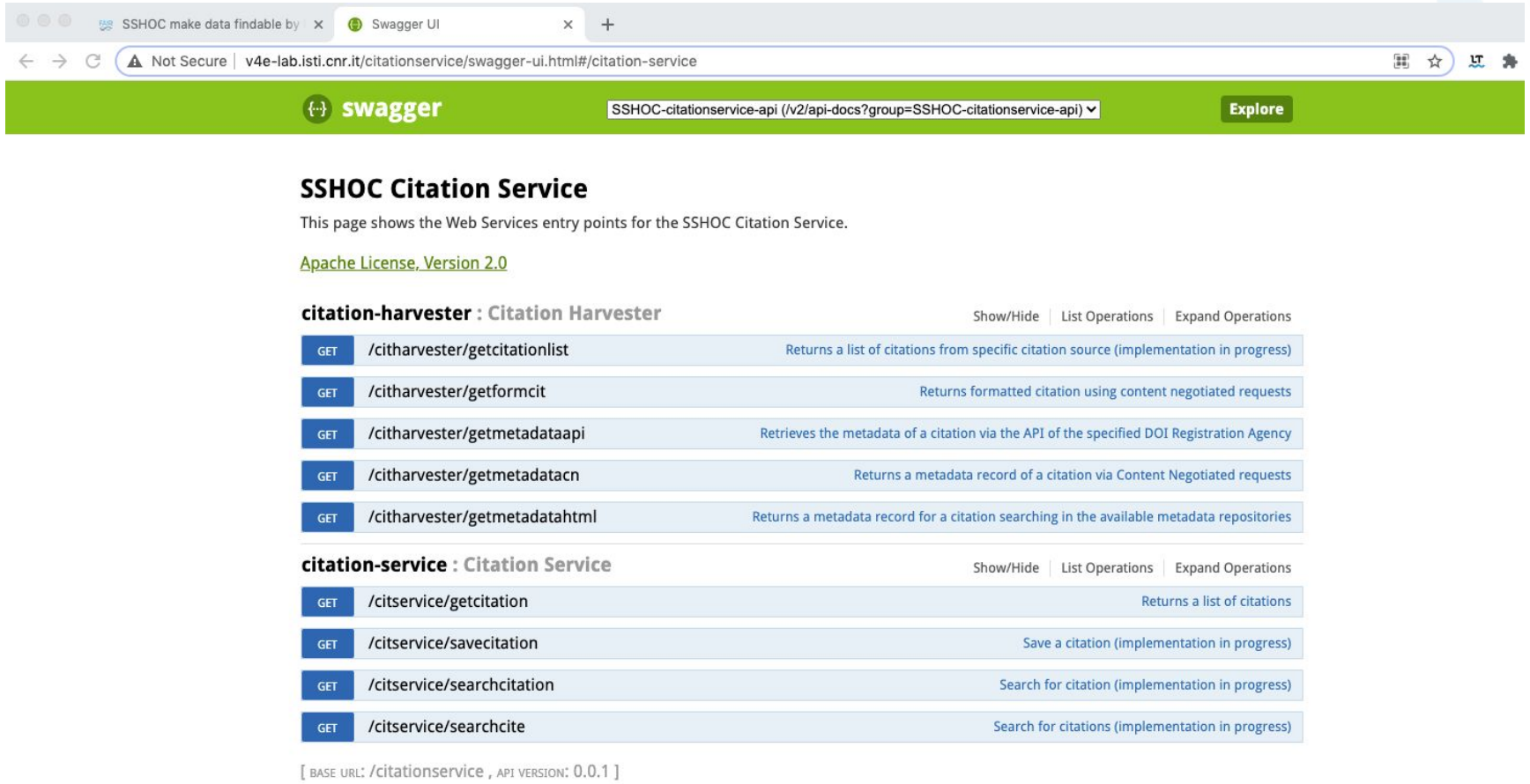
The SSHOC Citation Service Prototype



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

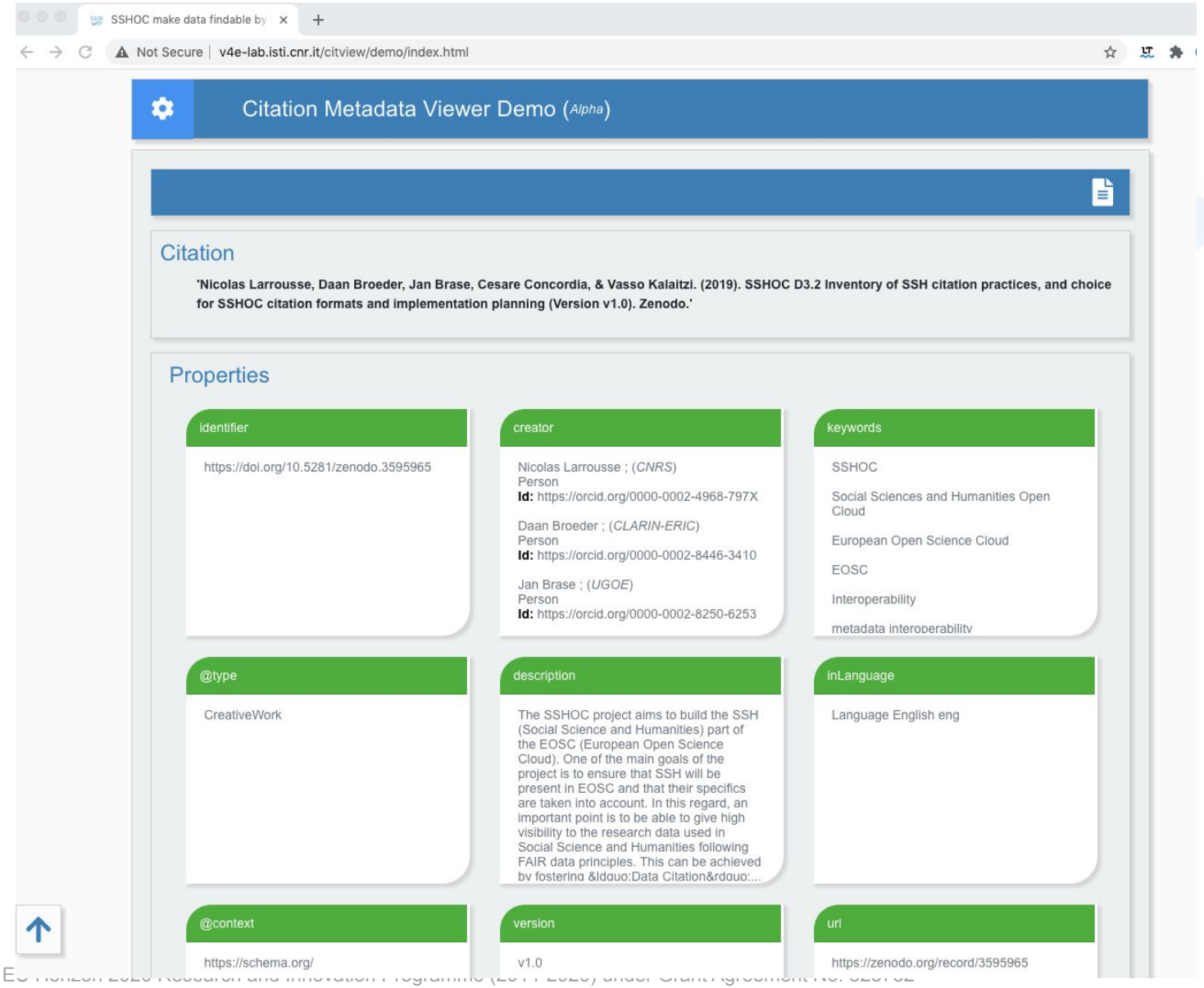
The Citation Service API



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

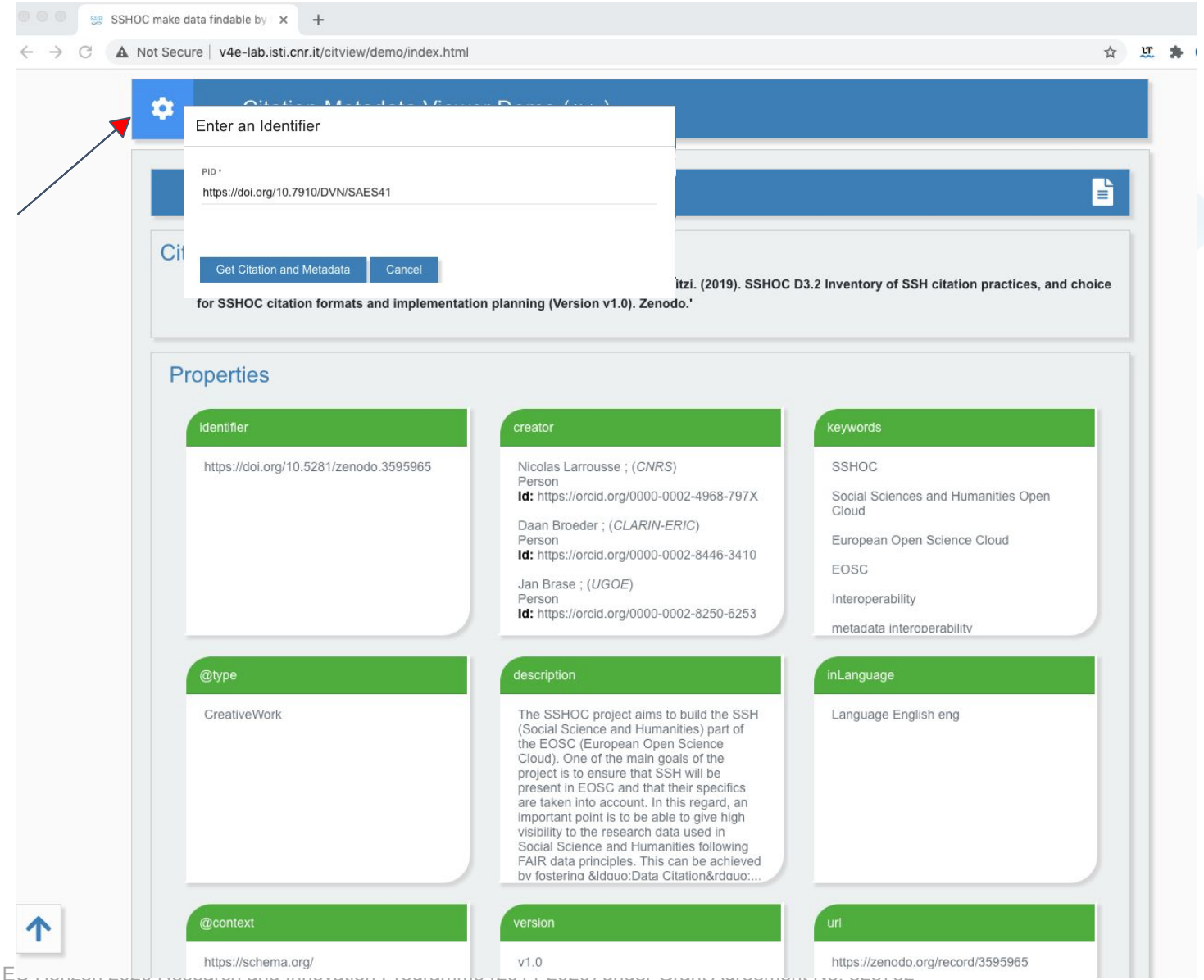
The Citation Metadata Viewer



This project is funded from the EC Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

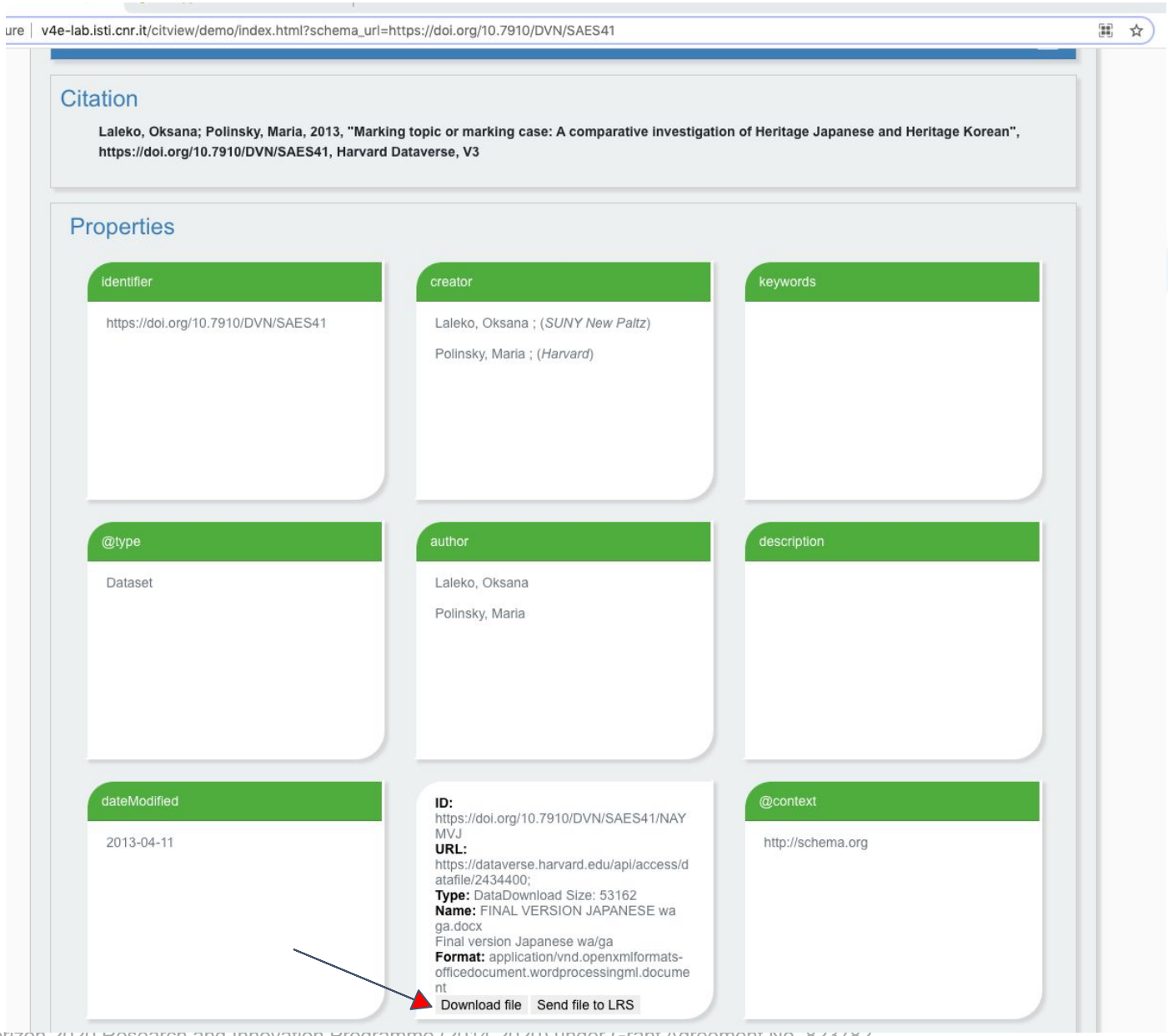
The Citation Metadata Viewer



This project is funded from the EC Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

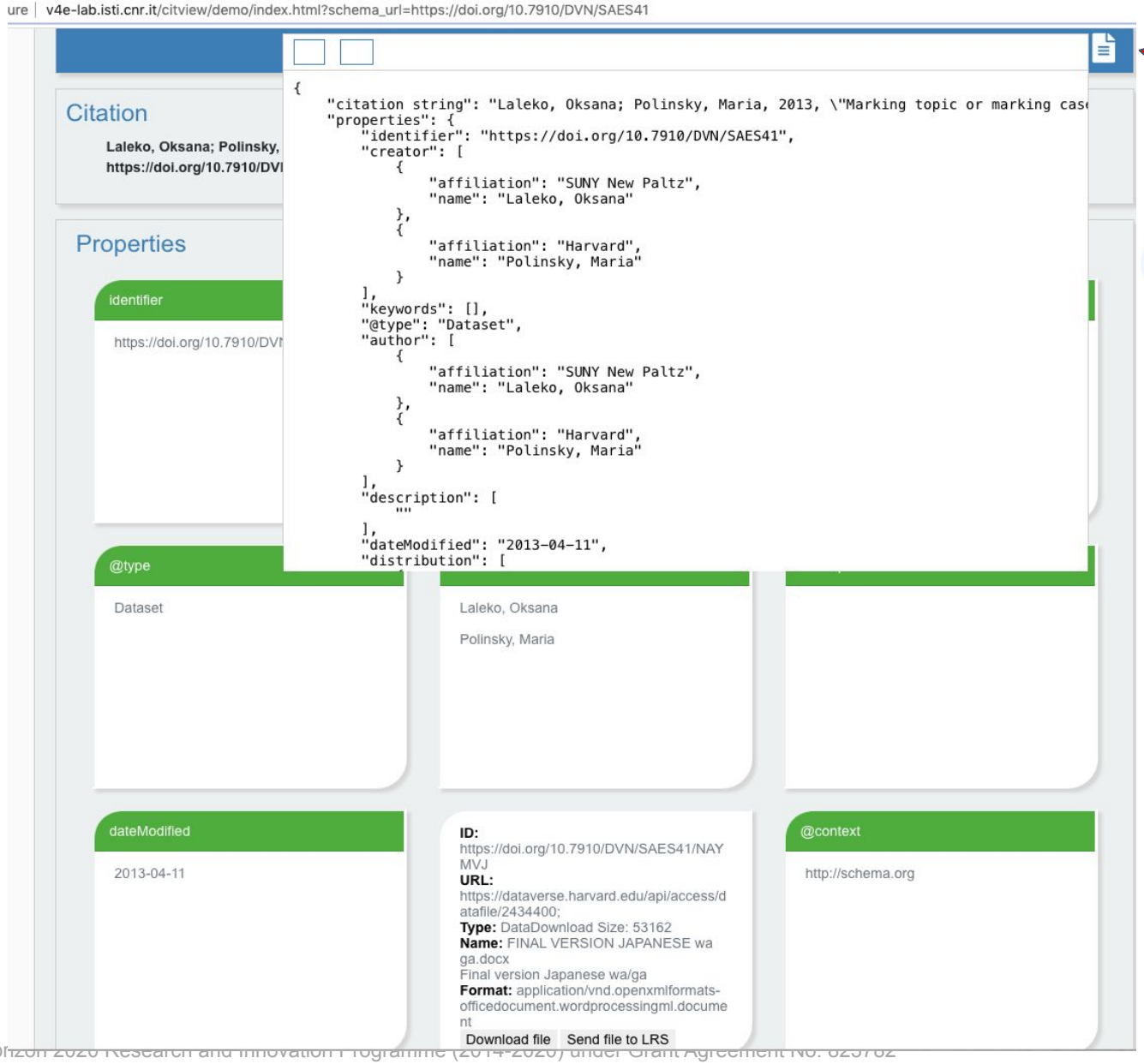
The Citation Metadata Viewer



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

The Citation Metadata Viewer



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

The SSHOC Citation Service Prototype

- Citation Service API:
  http://v4e-lab.isti.cnr.it/citationservice/swagger-ui.html#/

- The Citation Metadata Viewer:
  http://v4e-lab.isti.cnr.it/citview/demo/index.html

- Checking citations from the abstracts of all (ADHO) DH conferences from 2015 to 2020 and from DHQ journal articles
  https://github.com/cesareconcordia/sshoc-resources/blob/master/CItationDHres-Check.ipynb

- ***Warning: the URLs above refer a development version of the prototype API and Viewer, will be changed in the future***

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Conclusion

Citation alone are useless … there is a need for a complete ecosystem

- Good documentation
- Norms and standards
- Trusted repositories
- Dissemination tool

Information about datasets are not available at the same place …

What's next

- From citations to data papers
- Using information coming from citations to associate tools to data

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

Any Questions?

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Thank you for your attention!

Please share your thoughts about the event:
https://forms.gle/9VMN99YizicUG6RR8

www.sshopencloud.eu

@SSHOpenCloud

/in/sshopencloud

info@sshopencloud.eu

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud