B

Data citation for the Humanities and Social Sciences: a special case?

Barbara McGillivray (University of Cambridge and The Alan Turing Institute)
Nicolas Larrousse (CNRS / Huma-Num - SSHOC Project WP3)
Daan Broeder (CLARIN ERIC - SSHOC Project WP3)

D

Open Science and Open Knowledge

- Our common goal is **Open Science,** meaning we promote openness, integrity, and reproducibility in research;
- and the results to be treated as **Open Knowledge**; knowledge that is free to use, reuse, and redistribute without legal, social or technological restriction
- The **FAIR principles** help us with both
  - *Findability*, Accessibility, Interoperability, *Reusability*
- We (researchers, librarians etc.) like it, public funders like and require it

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: 48
(23 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

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

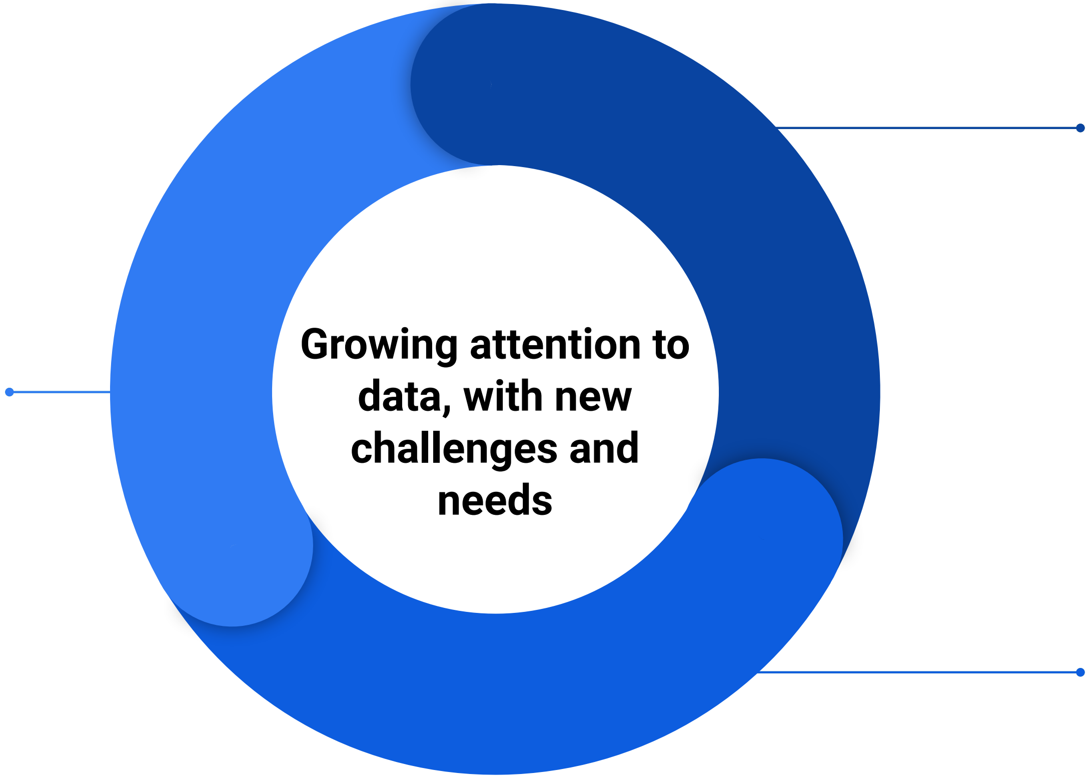
Project budget:
€ 14,455,594.08

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

Specificities of SSH



Growing adoption of data-intensive methods

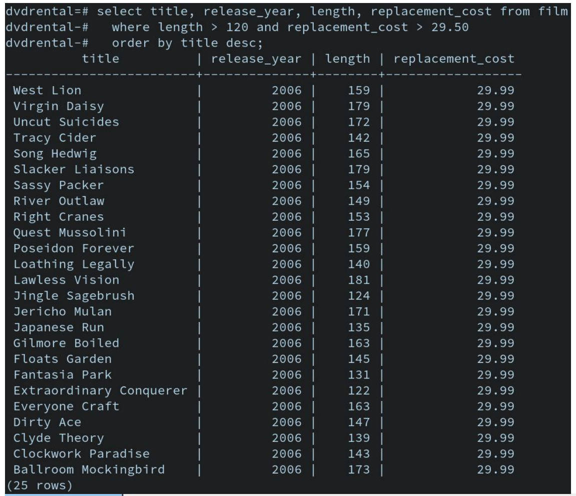
New types of digitized and born-digital data, and very diverse types of data, including cultural heritage

Growing attention to data, with new challenges and needs

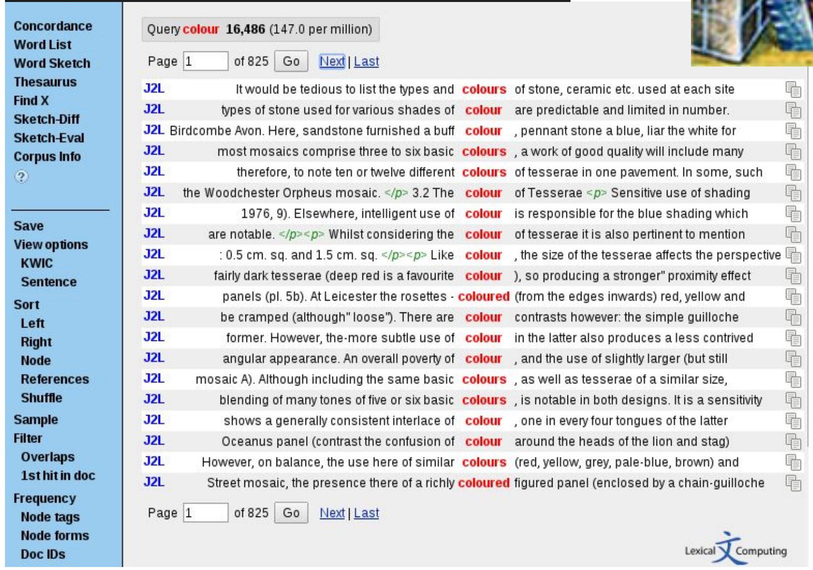
Increased number of data professionals and stewards in libraries

B

A variety of data types



```
dvdrental=# select title, release_year, length, replacement_cost from film
dvdrental-#   where length > 120 and replacement_cost > 29.50
dvdrental-#   order by title desc;
         title          | release_year | length | replacement_cost
------------------------+--------------+--------+------------------
 West Lion              |         2006 |    159 |            29.99
 Virgin Daisy           |         2006 |    179 |            29.99
 Uncut Suicides         |         2006 |    172 |            29.99
 Tracy Cider            |         2006 |    142 |            29.99
 Song Hedwig            |         2006 |    165 |            29.99
 Slacker Liaisons       |         2006 |    179 |            29.99
 Sassy Packer           |         2006 |    154 |            29.99
 River Outlaw           |         2006 |    149 |            29.99
 Right Cranes           |         2006 |    153 |            29.99
 Quest Mussolini        |         2006 |    177 |            29.99
 Poseidon Forever       |         2006 |    159 |            29.99
 Loathing Legally       |         2006 |    140 |            29.99
 Lawless Vision         |         2006 |    181 |            29.99
 Jingle Sagebrush       |         2006 |    124 |            29.99
 Jericho Mulan          |         2006 |    171 |            29.99
 Japanese Run           |         2006 |    135 |            29.99
 Gilmore Boiled         |         2006 |    163 |            29.99
 Floats Garden          |         2006 |    145 |            29.99
 Fantasia Park          |         2006 |    131 |            29.99
 Extraordinary Conquerer |        2006 |    122 |            29.99
 Everyone Craft         |         2006 |    163 |            29.99
 Dirty Ace              |         2006 |    147 |            29.99
 Clyde Theory           |         2006 |    139 |            29.99
 Clockwork Paradise     |         2006 |    143 |            29.99
 Ballroom Mockingbird   |         2006 |    173 |            29.99
(25 rows)
```

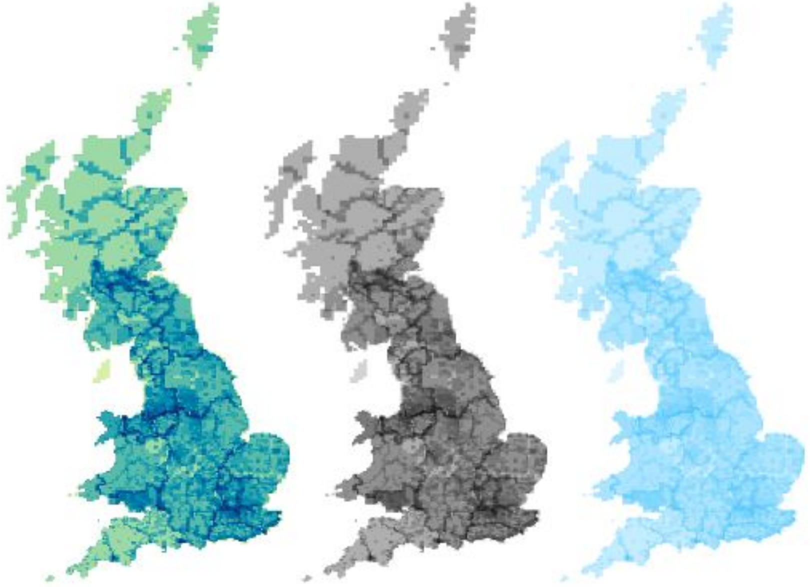


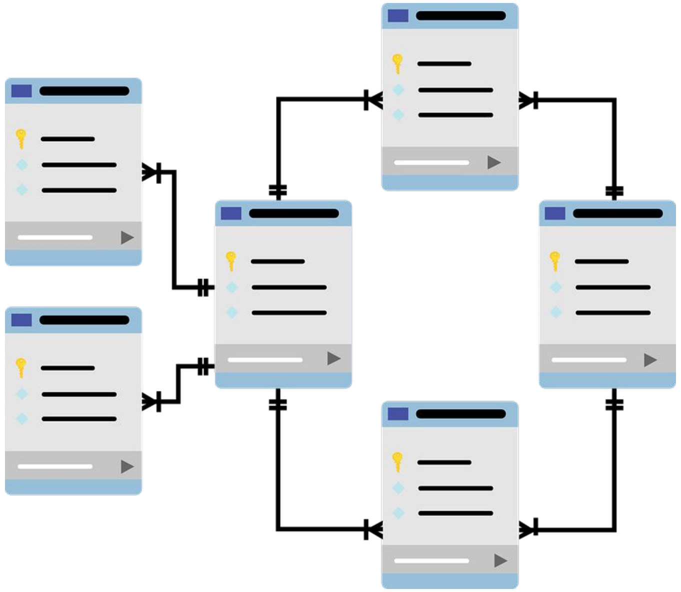
Query colour 16,486 (147.0 per million)

Page 1 of 825

J2L	It would be tedious to list the types and	colours	of stone, ceramic etc. used at each site
J2L	types of stone used for various shades of	colour	are predictable and limited in number.
J2L	Birdcombe Avon. Here, sandstone furnished a buff	colour	, pennant stone a blue, liar the white for
J2L	most mosaics comprise three to six basic	colours	, a work of good quality will include many
J2L	therefore, to note ten or twelve different	colours	of tesserae in one pavement. In some, such
J2L	the Woodchester Orpheus mosaic. </p> 3.2 The	colour	of Tesserae <p> Sensitive use of shading
J2L	1976, 9). Elsewhere, intelligent use of	colour	is responsible for the blue shading which
J2L	are notable. </p><p> Whilst considering the	colour	of tesserae it is also pertinent to mention
J2L	: 0.5 cm. sq. and 1.5 cm. sq. </p><p> Like	colour	, the size of the tesserae affects the perspective
J2L	fairly dark tesserae (deep red is a favourite	colour	), so producing a stronger" proximity effect
J2L	panels (pl. 5b). At Leicester the rosettes -	coloured	(from the edges inwards) red, yellow and
J2L	be cramped (although" loose"). There are	colour	contrasts however: the simple guilloche
J2L	former. However, the-more subtle use of	colour	in the latter also produces a less contrived
J2L	angular appearance. An overall poverty of	colour	, and the use of slightly larger (but still
J2L	mosaic A). Although including the same basic	colours	, as well as tesserae of a similar size,
J2L	blending of many tones of five or six basic	colours	, is notable in both designs. It is a sensitivity
J2L	shows a generally consistent interlace of	colour	, one in every four tongues of the latter
J2L	Oceanus panel (contrast the confusion of	colour	around the heads of the lion and stag)
J2L	However, on balance, the use here of similar	colours	(red, yellow, grey, pale-blue, brown) and
J2L	Street mosaic, the presence there of a richly	coloured	figured panel (enclosed by a chain-guilloche

Page 1 of 825





B

Data traditions and data practices

Variety

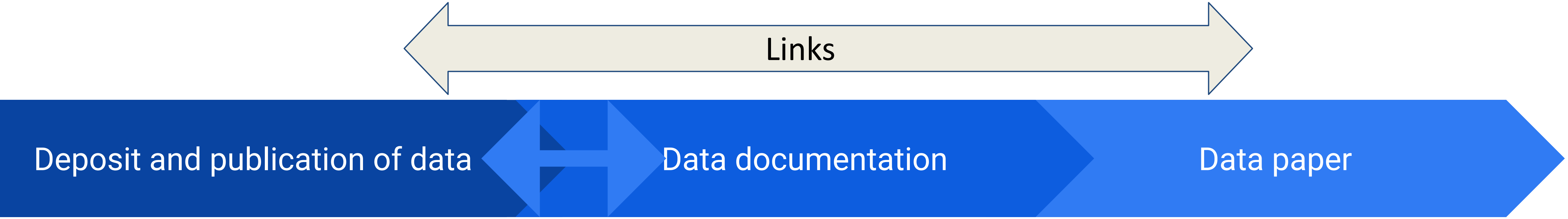
Different traditions for quantitative/computational methods

Focus

Historically, on publications (book) rather than data

B

From data to data papers



Links

Deposit and publication of data

Data documentation

Data paper

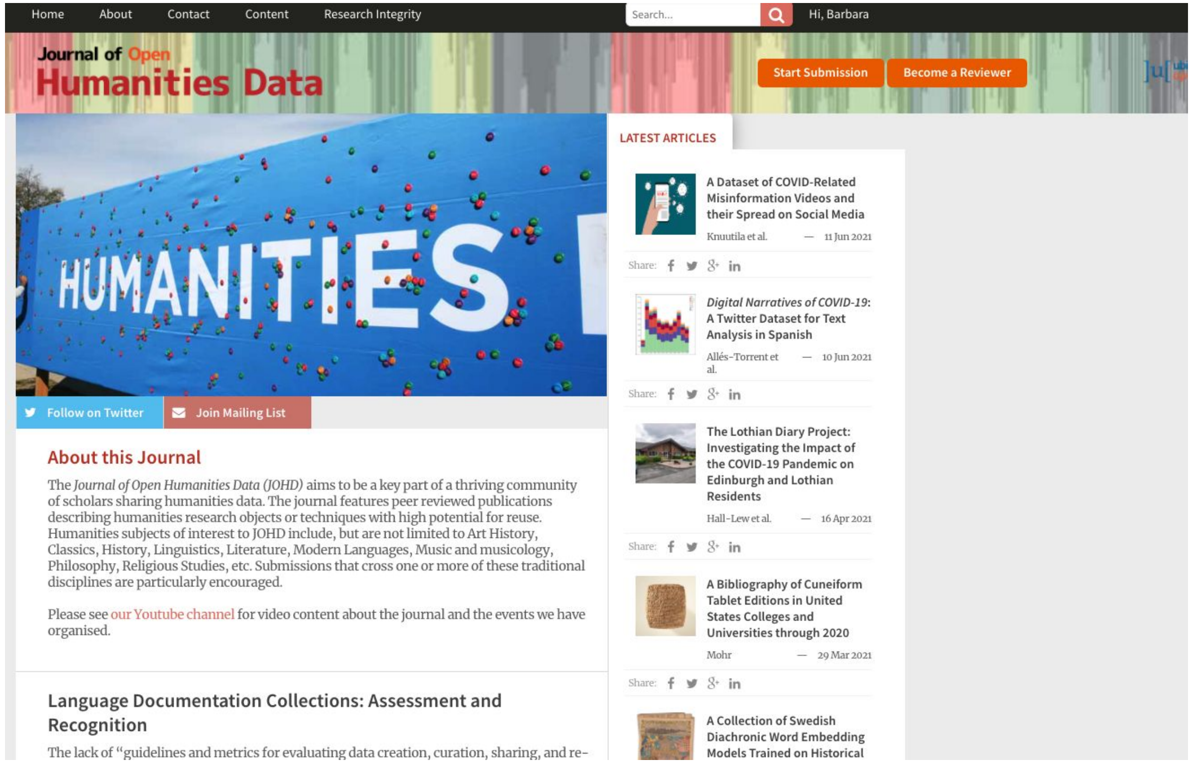
Typically in open repositories assigning DOI

+intellectual organisation

Peer-reviewed publications describing datasets and crediting their creators

The Journal of Open Humanities Data

- Launched in 2015; open access and peer-reviewed
- Vision: be a key part of a thriving community of scholars sharing humanities data
- Fastest growing data journal for humanities research
- 25 articles published so far (6 in 2021)
- Two publication types
  - Short data papers (1000 words): descriptions of datasets with high reuse potential
  - Full-length research papers (3000-5000 words): discussion of methods, challenges, and limitations in the creation, collection, management of humanities data



About this Journal

The *Journal of Open Humanities Data (JOHD)* aims to be a key part of a thriving community of scholars sharing humanities data. The journal features peer reviewed publications describing humanities research objects or techniques with high potential for reuse. Humanities subjects of interest to JOHD include, but are not limited to Art History, Classics, History, Linguistics, Literature, Modern Languages, Music and musicology, Philosophy, Religious Studies, etc. Submissions that cross one or more of these traditional disciplines are particularly encouraged.

Please see our Youtube channel for video content about the journal and the events we have organised.

Language Documentation Collections: Assessment and Recognition

The lack of "guidelines and metrics for evaluating data creation, curation, sharing, and re-

A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media
Knuutila et al. — 11 Jun 2021

Digital Narratives of COVID-19: A Twitter Dataset for Text Analysis in Spanish
Allés-Torrent et al. — 10 Jun 2021

The Lothian Diary Project: Investigating the Impact of the COVID-19 Pandemic on Edinburgh and Lothian Residents
Hall-Lew et al. — 16 Apr 2021

A Bibliography of Cuneiform Tablet Editions in United States Colleges and Universities through 2020
Mohr — 29 Mar 2021

A Collection of Swedish Diachronic Word Embedding Models Trained on Historical

https://openhumanitiesdata.metajnl.com/

Data citation in SSH

- The notion of “publishing data” is relatively new
- Data until a recent period was not really “*noble*” … the focus was more on the “final product” (e.g. an article or a book)

- Very diverse and no specific common approach to data citation A lot of initiatives … so far no real standard, but more “communities of practices” (See SSHOC deliverable ” Inventory of SSH citation practices ….” https://doi.org/10.5281/zenodo.4436736)

REVUE D’ÉGYPTOLOGIE 67

PUBLIÉE PAR LA SOCIÉTÉ FRANÇAISE D’ÉGYPTOLOGIE

PARIS – ÉDITIONS PEETERS

2016

-> In order to cite data, you need **infrastructures** to publish them in a proper (FAIR) way but you also must develop mechanism(s) to cite SSH data:

- Build stronger links between data and publication
- Facilitate and develop a “culture of data publication/citation”

etc.

Digital Medievalist

SSHOC
social sciences & humanities open cloud

N

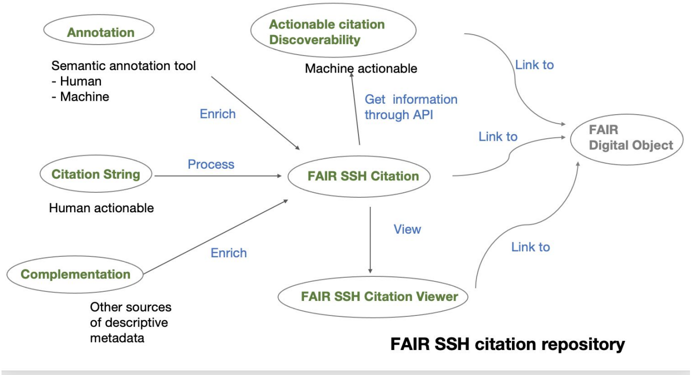
One goal of SSHOC is to address the problems of the current data landscape with its disciplinary silos

SSHOC Project and Data citation

- Recommendations based on “Force11 Citation Principles” adapted to SSH specificities
- Make citations “actionable” as a classic citation isn’t machine actionable

-> Proposition to create a prototype of *“FAIR SSH Citation Infrastructure”*

- Process existing citations or create new ones
- Enrich them automatically (other sources) and manually (annotations)
- Standardized citations
- Publish citations into the “Citation Infrastructure” to make them actionable



D

The future for data publication and citation

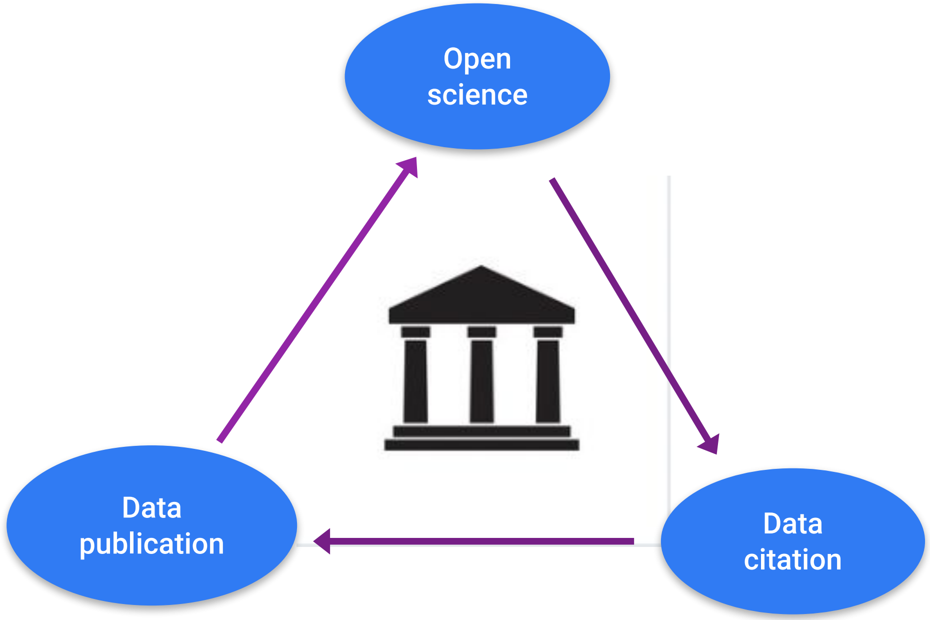
Although the potential for data creation, processing and (re)sharing have increased enormously

- Data citation practices still mainly based on traditional publication model ie. paper citing paper citing paper
- More complex collaborative and dynamic workflows are possible
- and also needed for using distributed and dynamic data sources e.g. social media and virtual collections of heterogeneous and distributed data
- Key is proper data identification, description and provenance tracking during the whole data life-cycle

Libraries in the centre?

- Increased need for Openness and Sharing of data and information at all phases of the SSH DLC
- More data and new types of data
  - how do we cite millions of data points?
  - how do we cite dynamic data e.g. social media
- Are the current practices wrt data publication and citation still valid?

- Libraries may ask if they can go far beyond caring only for the end-result (papers & data),
- they are already involved in
  - storing project result data-sets for reproducibility
  - training and facilitating researchers wrt Open Science and Open Knowledge and FAIR principles
  - Providing catalogues and meta-science information
- However the physical and conceptual distance between Libraries and research labs poses limitations


Open science
Data citation
Data publication