Rich metadata for annotation of contexts for citation and data-citation

C.M. Zwölf, N. Moreau, Y-.A Ba, M-.L. Dubernet
and VAMDC consortium

VAMDC consortium

LERMA l'Observatoire de Paris | PSL

PADC Paris Astronomical Data Centre

Sketching the context

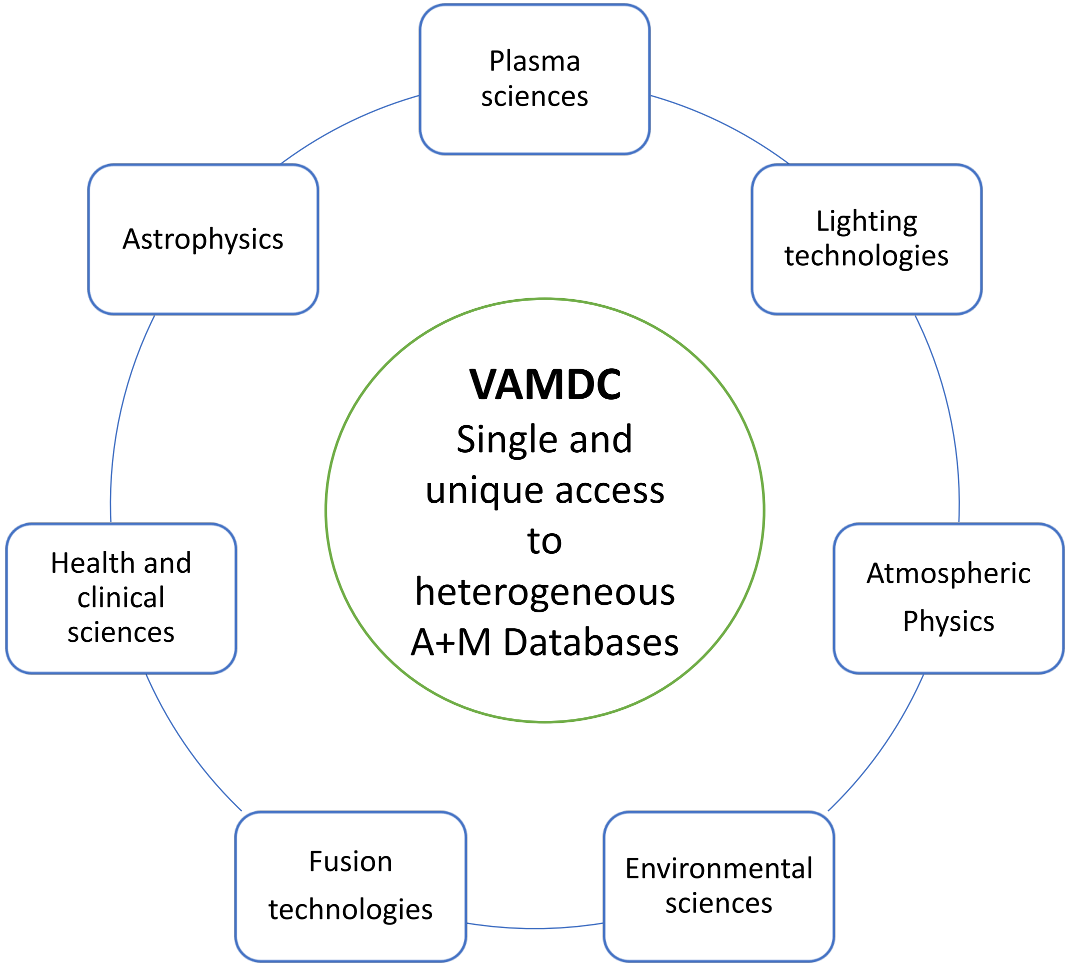
What is VAMDC

What are Data in VAMDC

Sketching the context

What is VAMDC

What are Data in VAMDC



- Worldwide interoperable e-infrastructure
- Federates ~30 heterogeneous databases http://portal.vamdc.org/
- The "V" of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- The consortium is politically organized around a Memorandum of understanding (15 international members have signed the MoU, 1 November 2014)
- High quality scientific data come from different Physical/Chemical Communities
- Provides data producers with a large dissemination platform
- Remove bottleneck between data-producers and wide body of users

Sketching the context

What is VAMDC

What are Data in VAMDC

Numerical quantities related to atomic and/or molecular process

Have been published into a scientific paper

Reviewed and assessed by the community

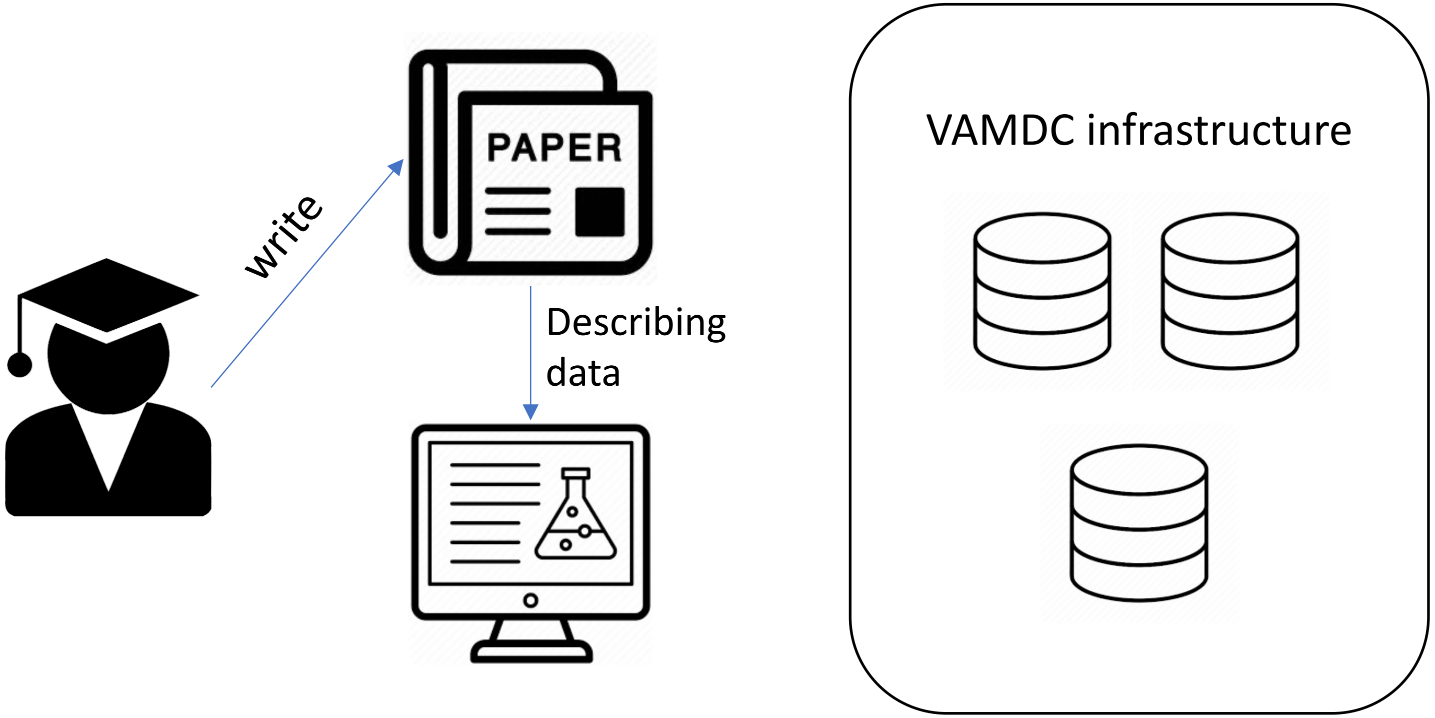
Entered into a VAMDC-federated database

Curated technically and scientifically by the database maintainer

A dual aspect of data-papers linking

Data Producers

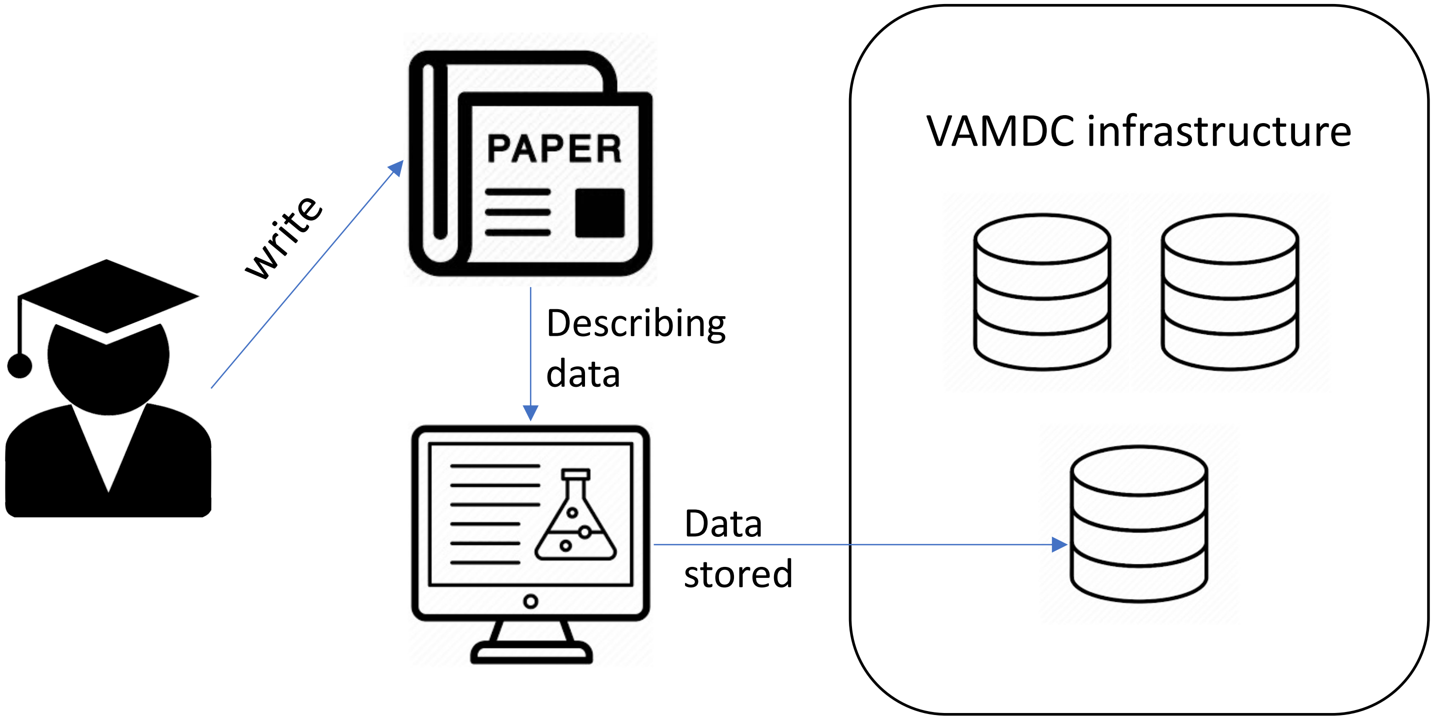
Data Users



A dual aspect of data-papers linking

Data Producers

Data Users



write

PAPER

Describing data

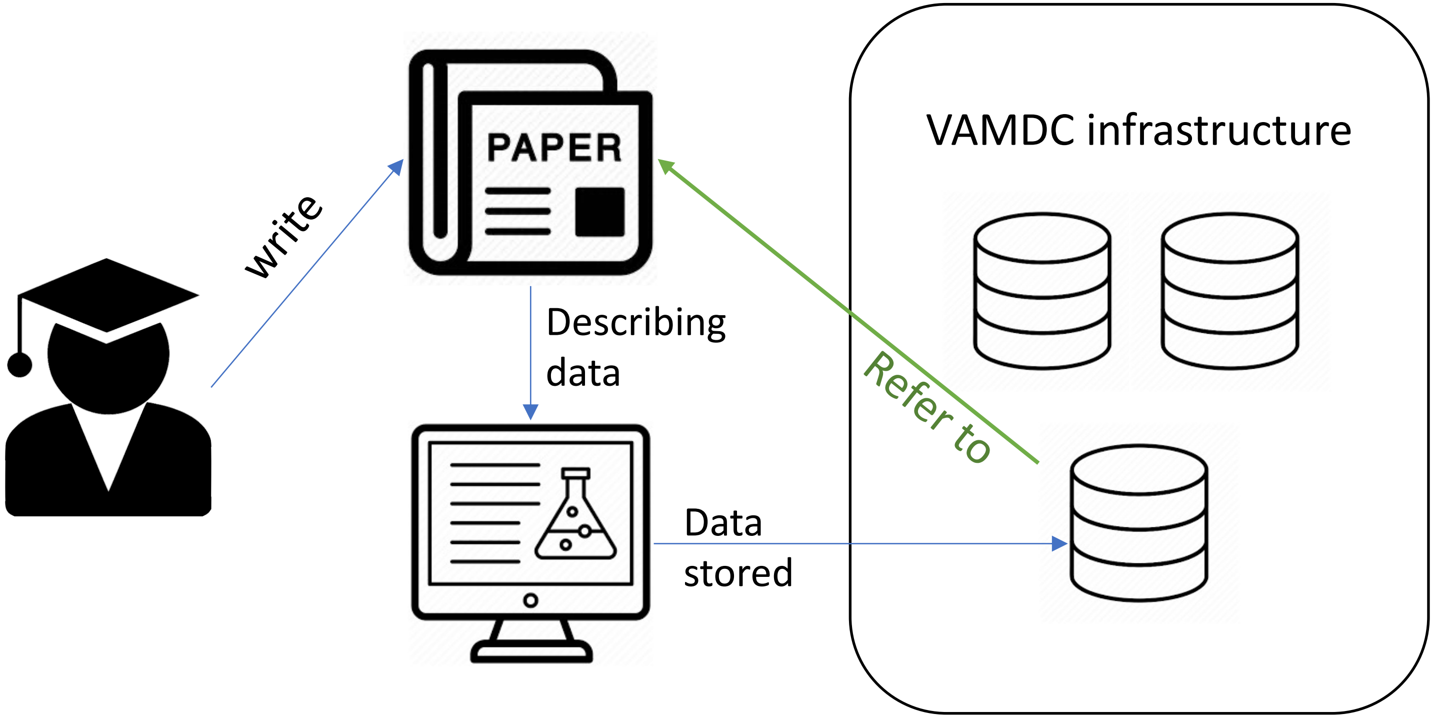
Data stored

VAMDC infrastructure

A dual aspect of data-papers linking

Data Producers

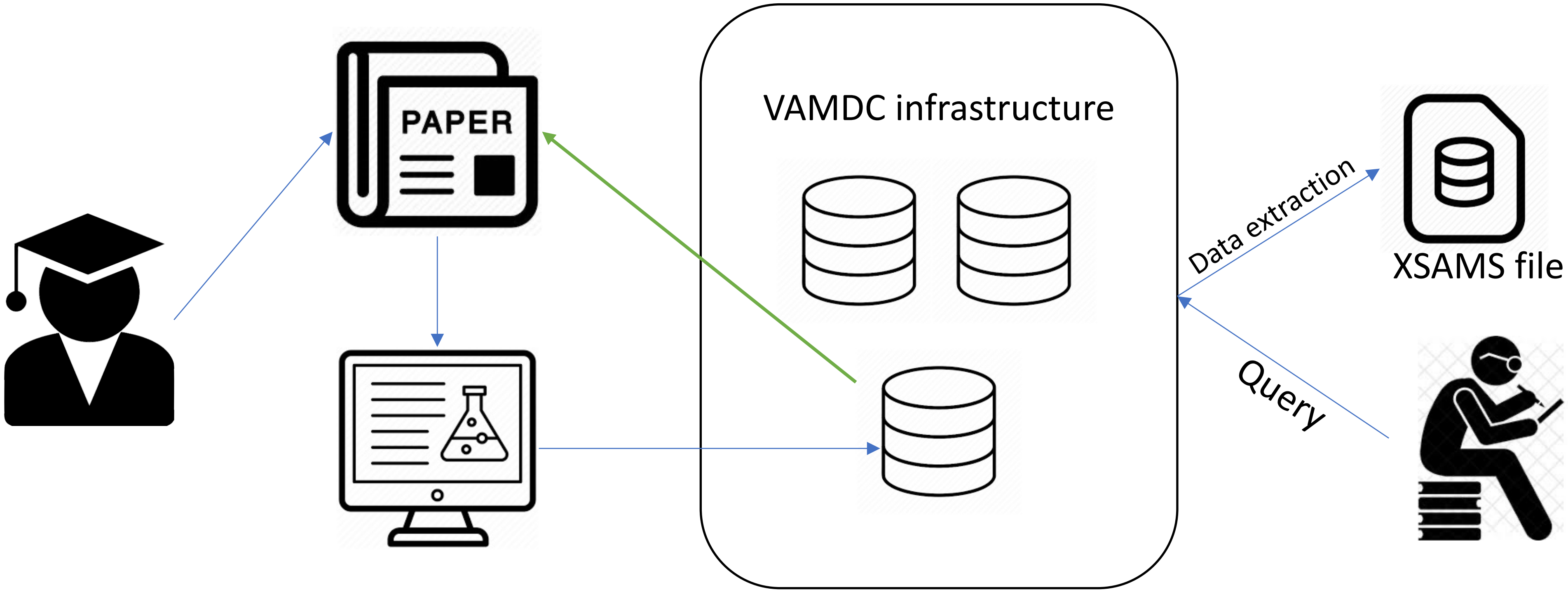
Data Users



A dual aspect of data-papers linking

Data Producers

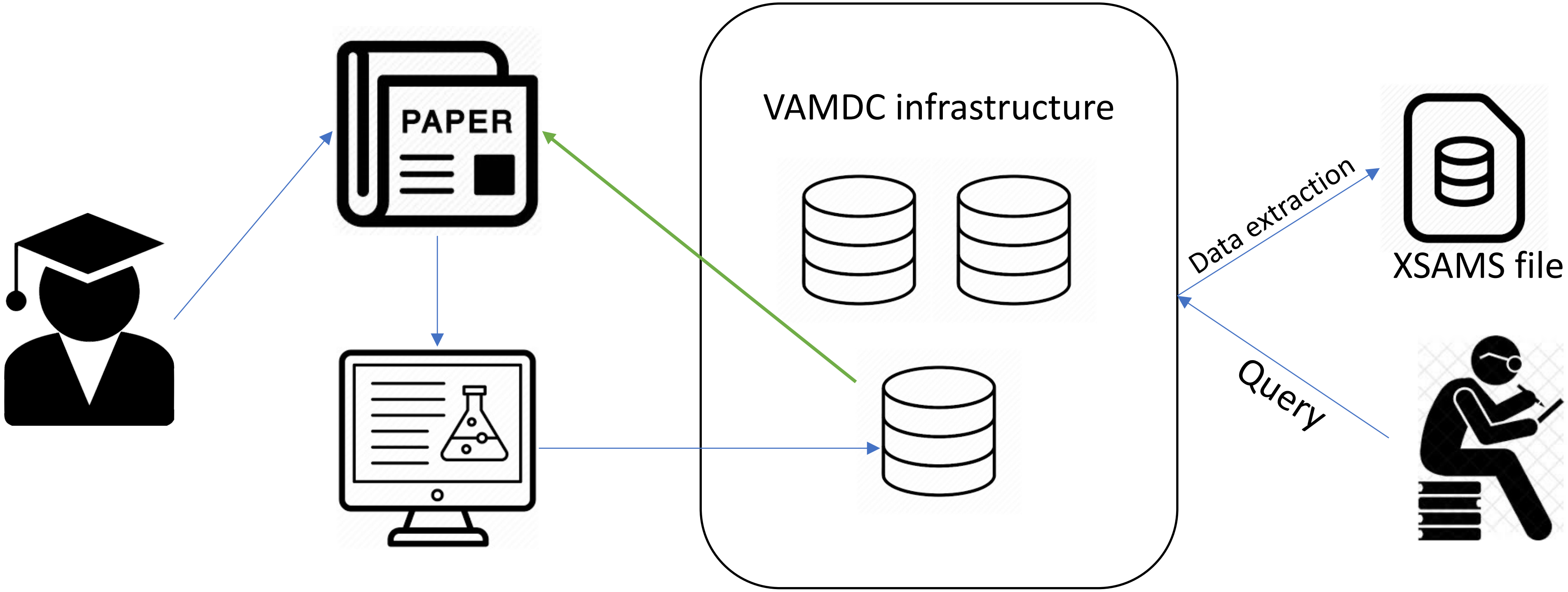
Data Users



A dual aspect of data-papers linking

Data Producers

Data Users

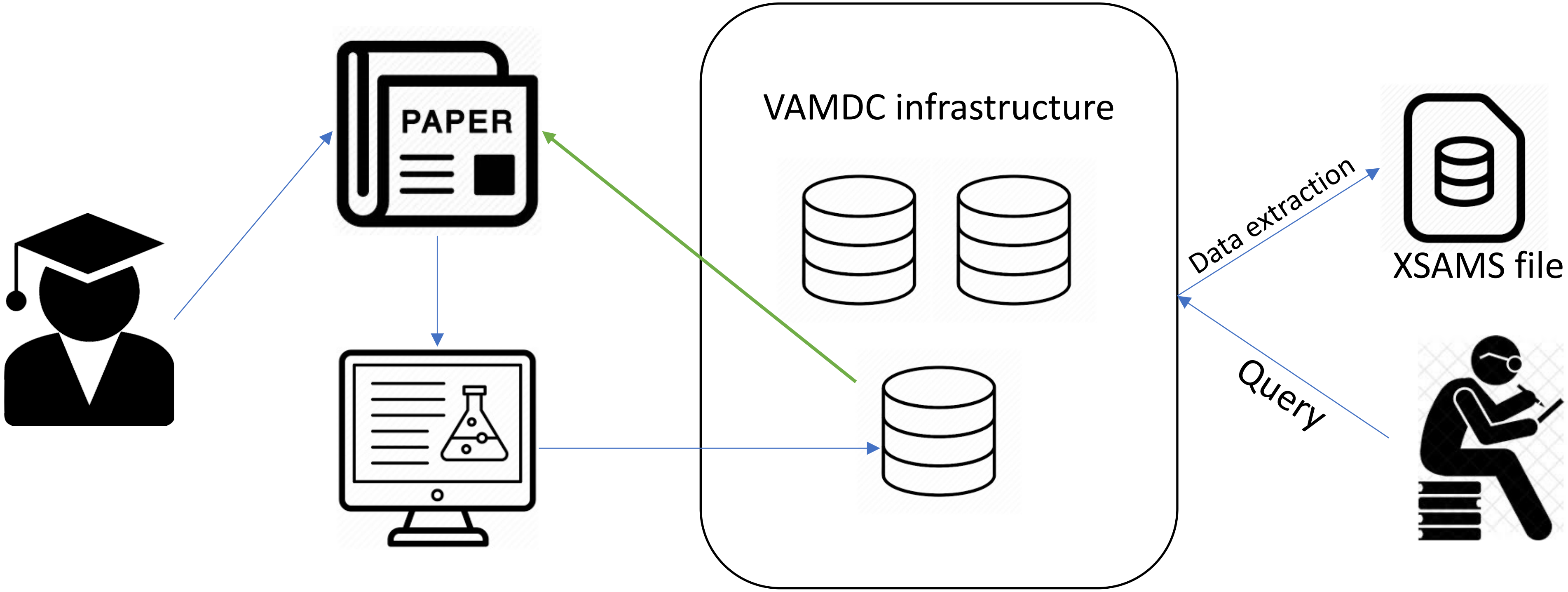


For further details, cf.
http://standards.vamdc.eu

A dual aspect of data-papers linking

Data Producers

Data Users

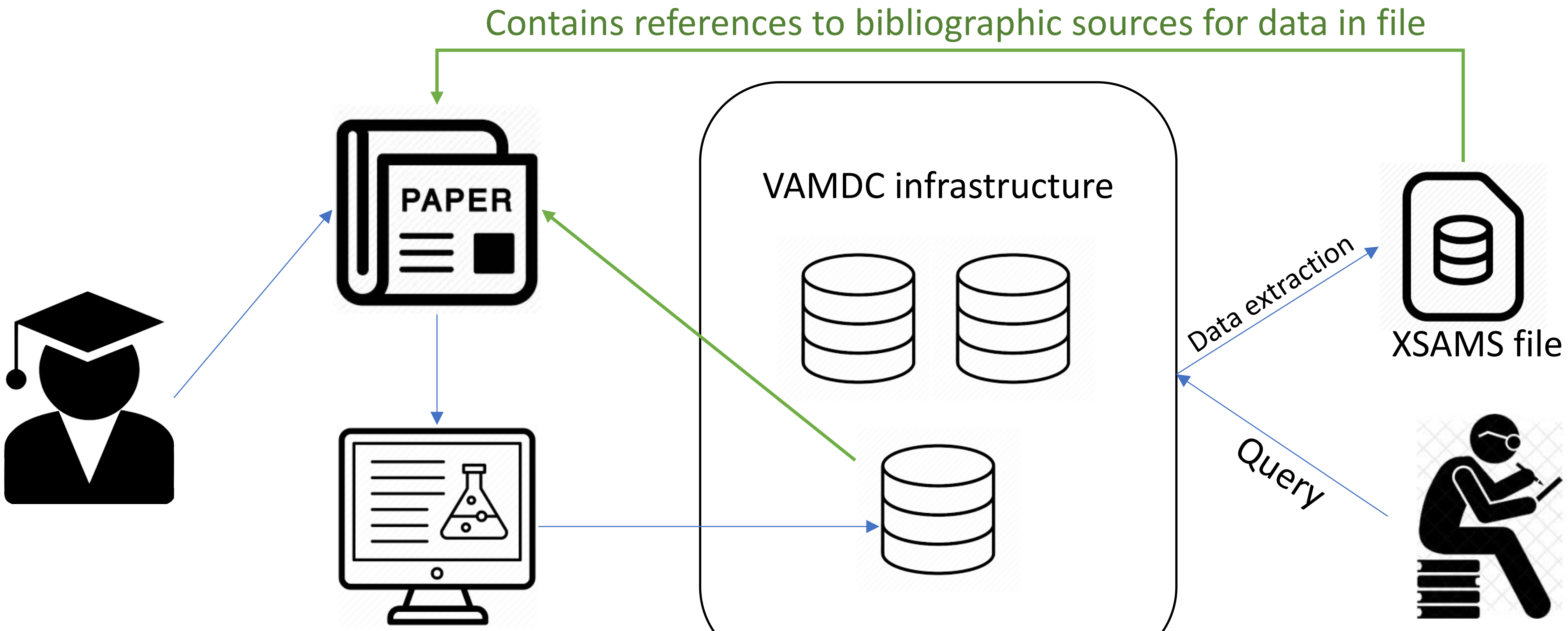


XSAMS is a rigorous and unambiguous object model for atomic and molecular physics: XML Schema for Atoms Molecules and Solids (joint effort from VAMDC, NIST, IAEA)

A dual aspect of data-papers linking

Data Producers

Data Users



Contains references to bibliographic sources for data in file

PAPER

VAMDC infrastructure

Data extraction

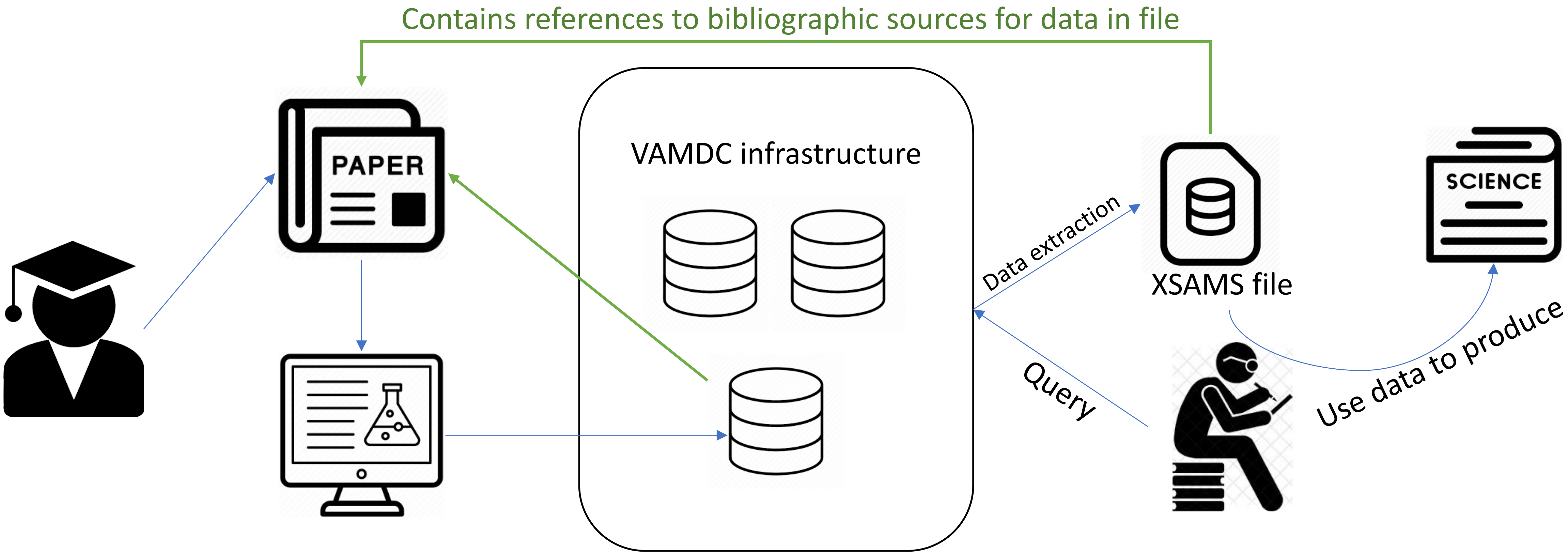
XSAMS file

Query

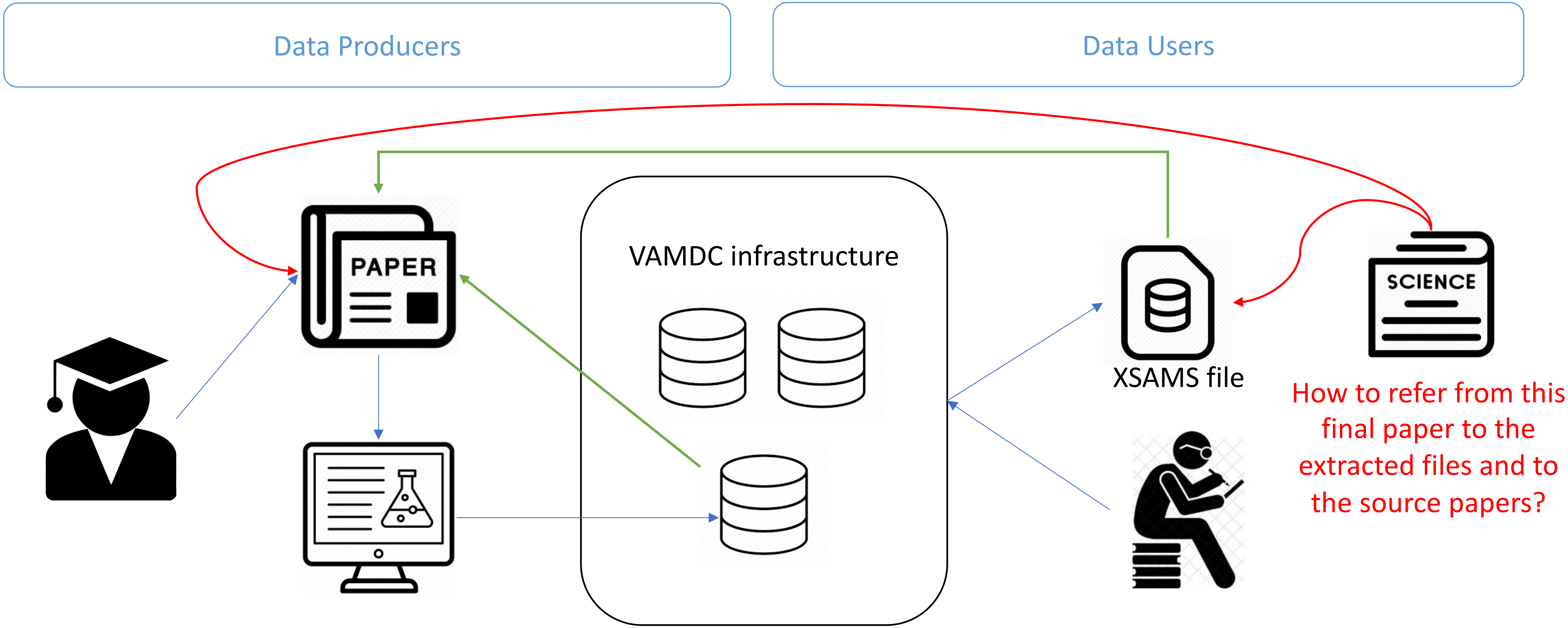
A dual aspect of data-papers linking

Data Producers

Data Users



A dual aspect of data-papers linking



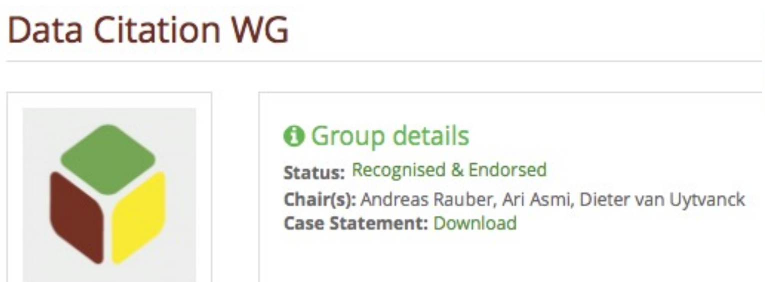
How to refer from this final paper to the extracted files and to the source papers?

A dual aspect of data-papers linking

To address this issue we started to work in 2014 with the Research Data Alliance

A dual aspect of data-papers linking

To address this issue we started to work in 2014 with the Research Data Alliance



Data Citation WG

Group details

Status: Recognised & Endorsed

Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck

Case Statement: Download

SCHOLIX

zenodo

Goals of this WG are to create identification mechanisms that:

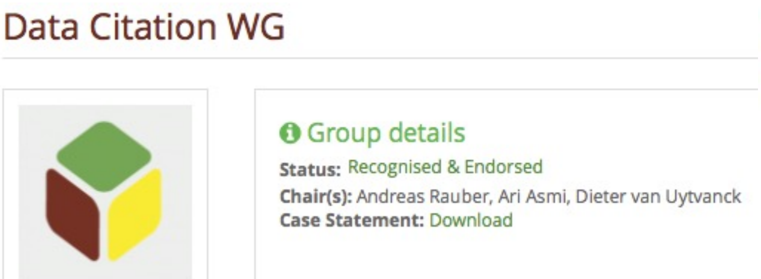
- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

A dual aspect of data-papers linking

To address this issue we started to work in 2014 with the Research Data Alliance



Data Citation WG

Group details

Status: Recognised & Endorsed

Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck

Case Statement: Download

SCHOLIX

zenodo

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

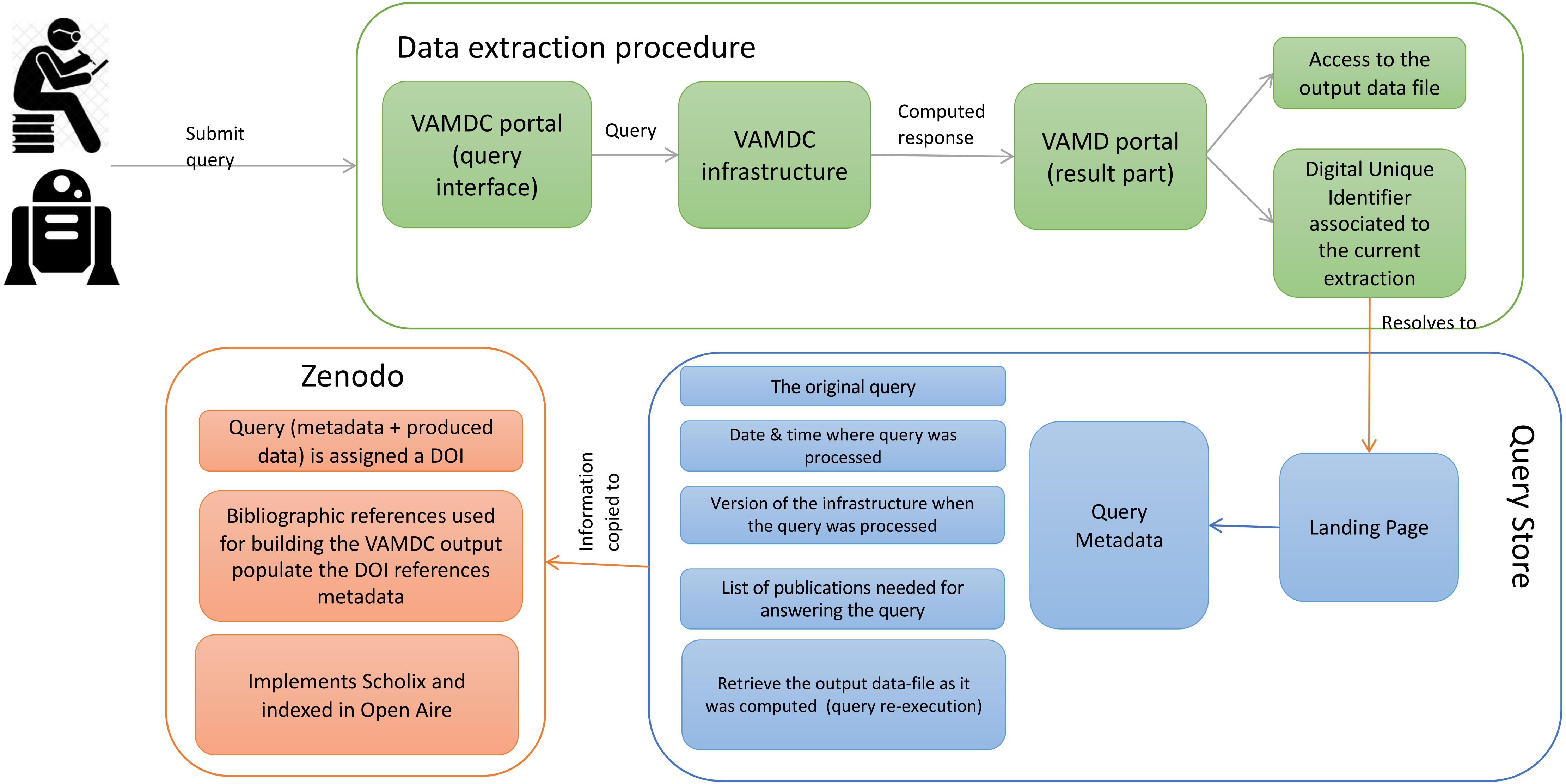
- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

Recommendation is to store all the queries (with their metadata) into a **Query Store (QS).**

The difficulty we had to cope with:

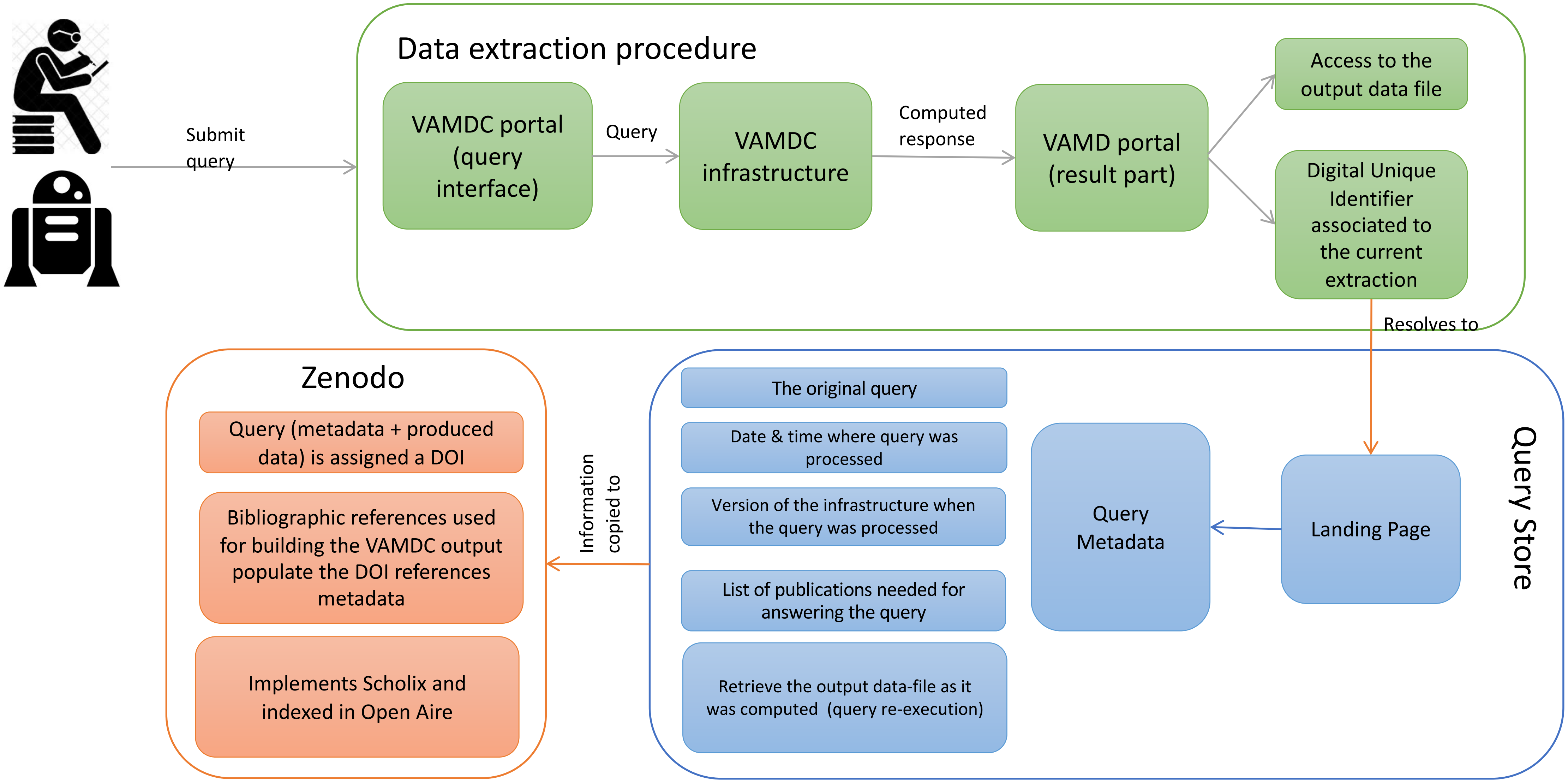
- How to handle a QS in the VAMDC distributed environment (VAMDC is a set of distributed services with no central management system)
- How to integrate the QS with the existing VAMDC components

The VAMDC Query Store


Submit query
Data extraction procedure
VAMDC portal (query interface)
Query
VAMDC infrastructure
Computed response
VAMD portal (result part)
Access to the output data file
Digital Unique Identifier associated to the current extraction
Resolves to
Zenodo
Query (metadata + produced data) is assigned a DOI
Bibliographic references used for building the VAMDC output populate the DOI references metadata
Implements Scholix and indexed in Open Aire
Information copied to
The original query
Date & time where query was processed
Version of the infrastructure when the query was processed
List of publications needed for answering the query
Retrieve the output data-file as it was computed (query re-execution)
Query Metadata
Landing Page
Query Store

- Data become directly citable by their DOI. Authors/papers referenced in the data-set will get credits automatically when the dataset is cited (using the DOI) into a paper
  - **Strong marketing argument: Put your data in VAMDC. You will get automatically credits each time your data is cited!**

The VAMDC Query Store



Submit query

Data extraction procedure

VAMDC portal (query interface)

Query

VAMDC infrastructure

Computed response

VAMD portal (result part)

Access to the output data file

Digital Unique Identifier associated to the current extraction

Resolves to

Query Store

Landing Page

Query Metadata

The original query

Date & time where query was processed

Version of the infrastructure when the query was processed

List of publications needed for answering the query

Retrieve the output data-file as it was computed (query re-execution)

Information copied to

Zenodo

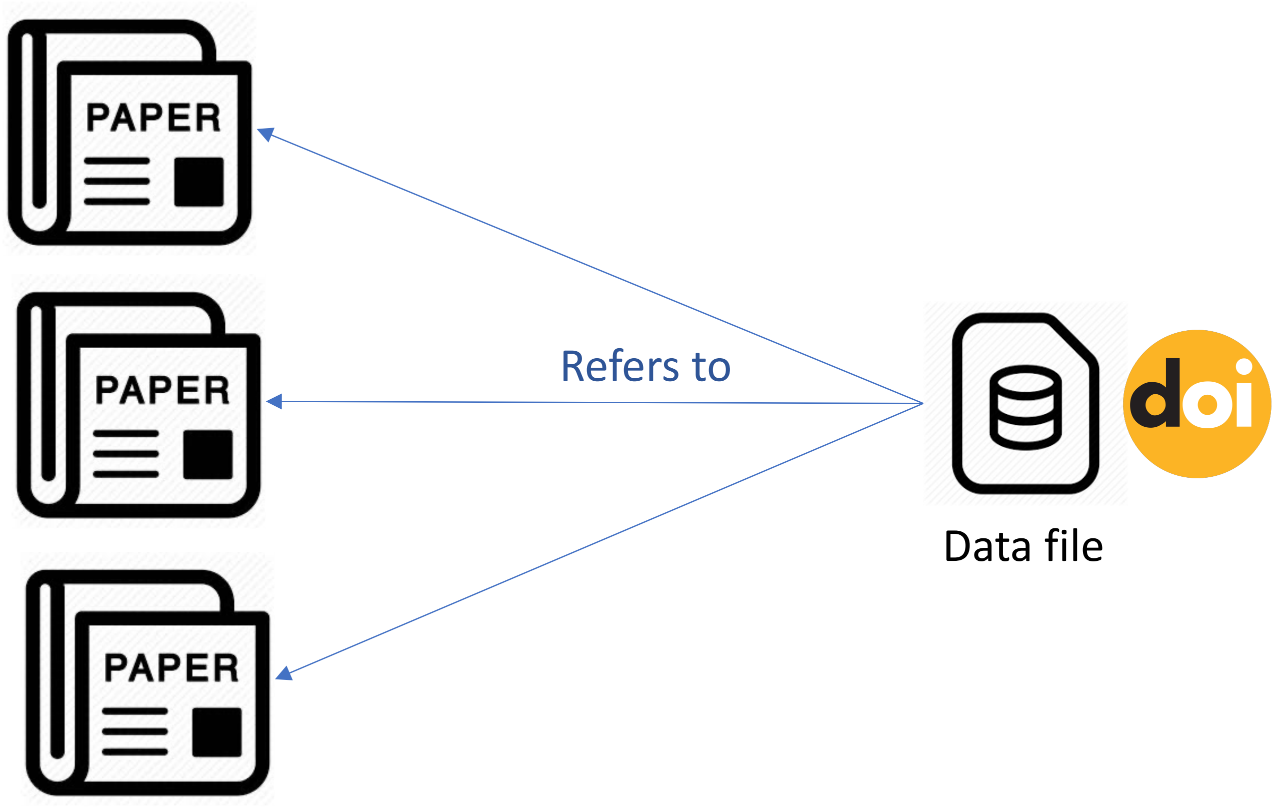
Query (metadata + produced data) is assigned a DOI

Bibliographic references used for building the VAMDC output populate the DOI references metadata

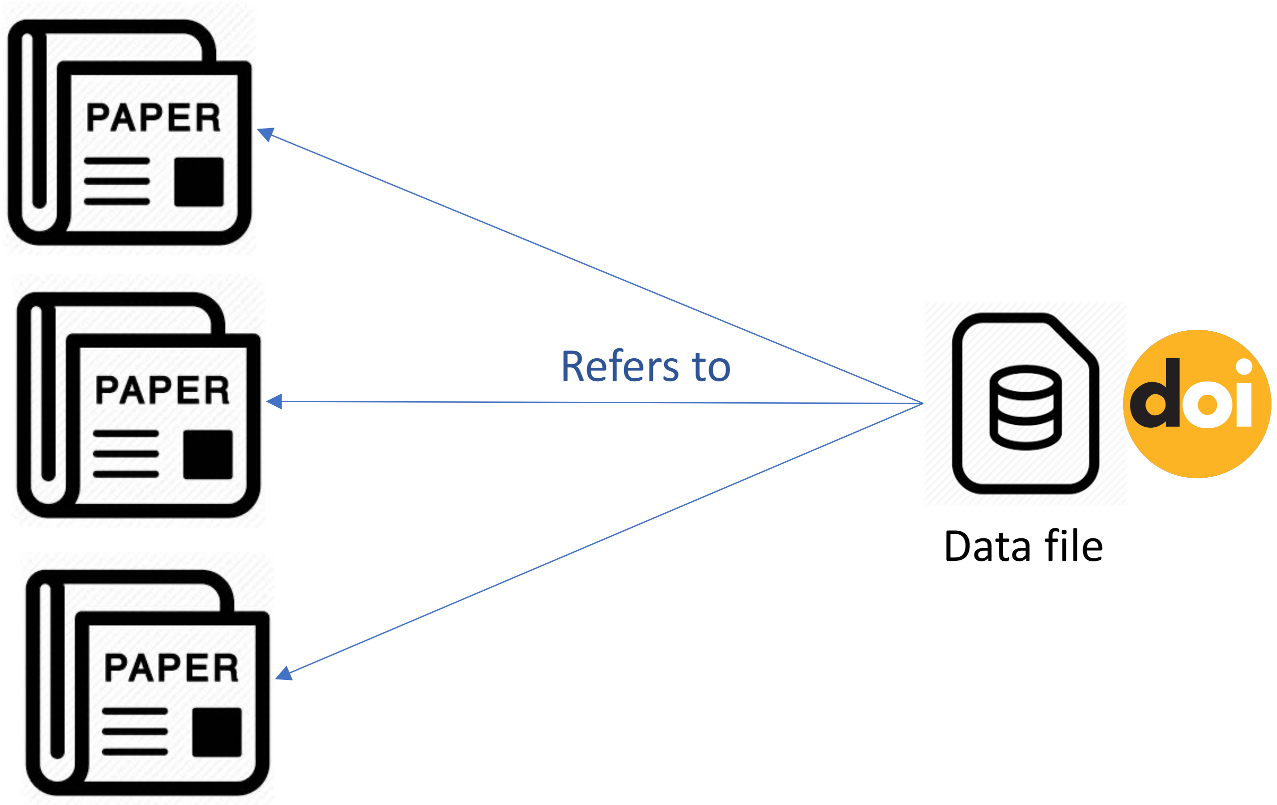
Implements Scholix and indexed in Open Aire

- Data become directly citable by their DOI. Authors/papers referenced in the data-set will get credits automatically when the dataset is cited (using the DOI) into a paper

The sources of a new challenge



The sources of a new challenge

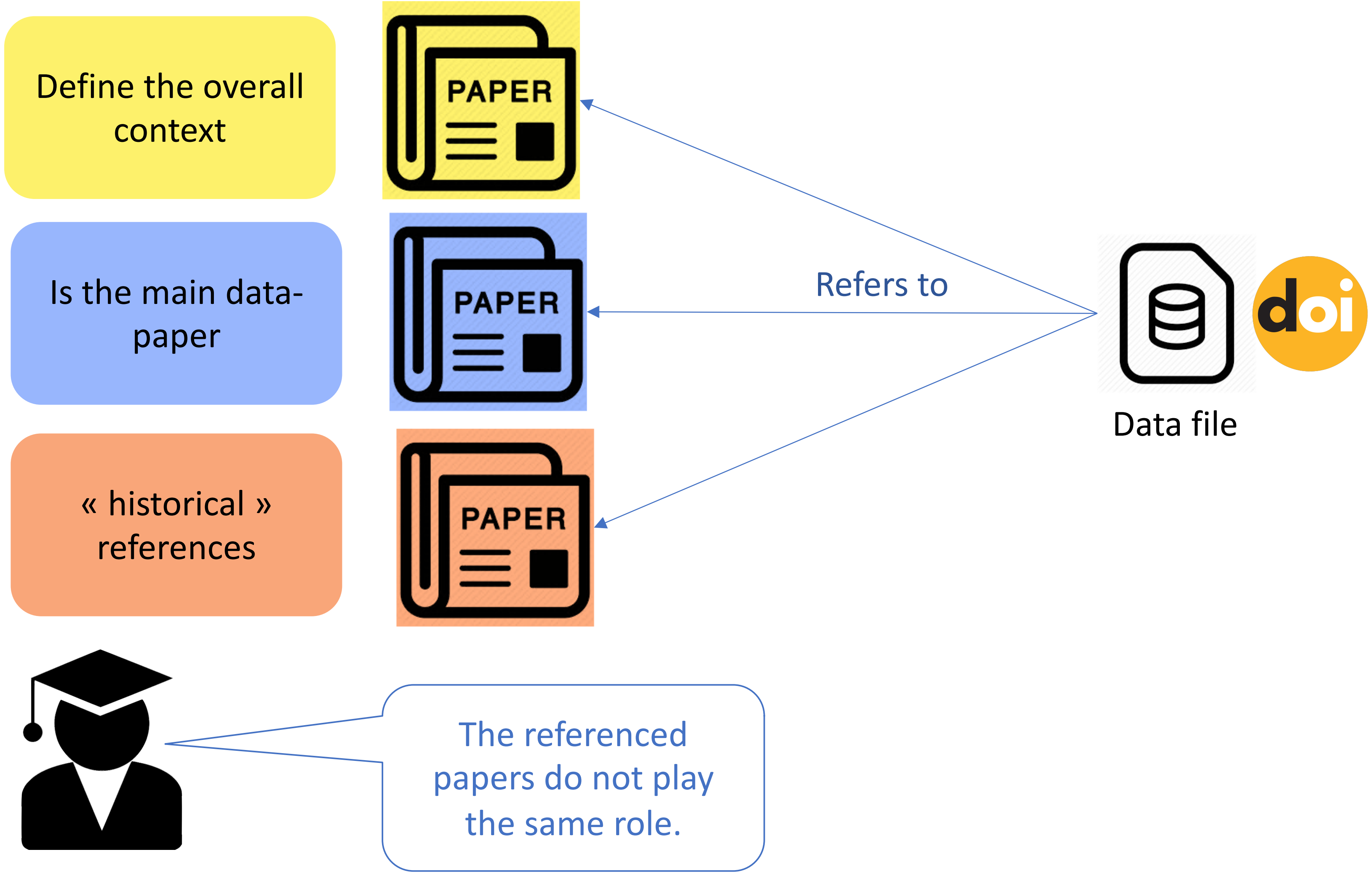


Refers to

Data file

The referenced papers do not play the same role.

The sources of a new challenge



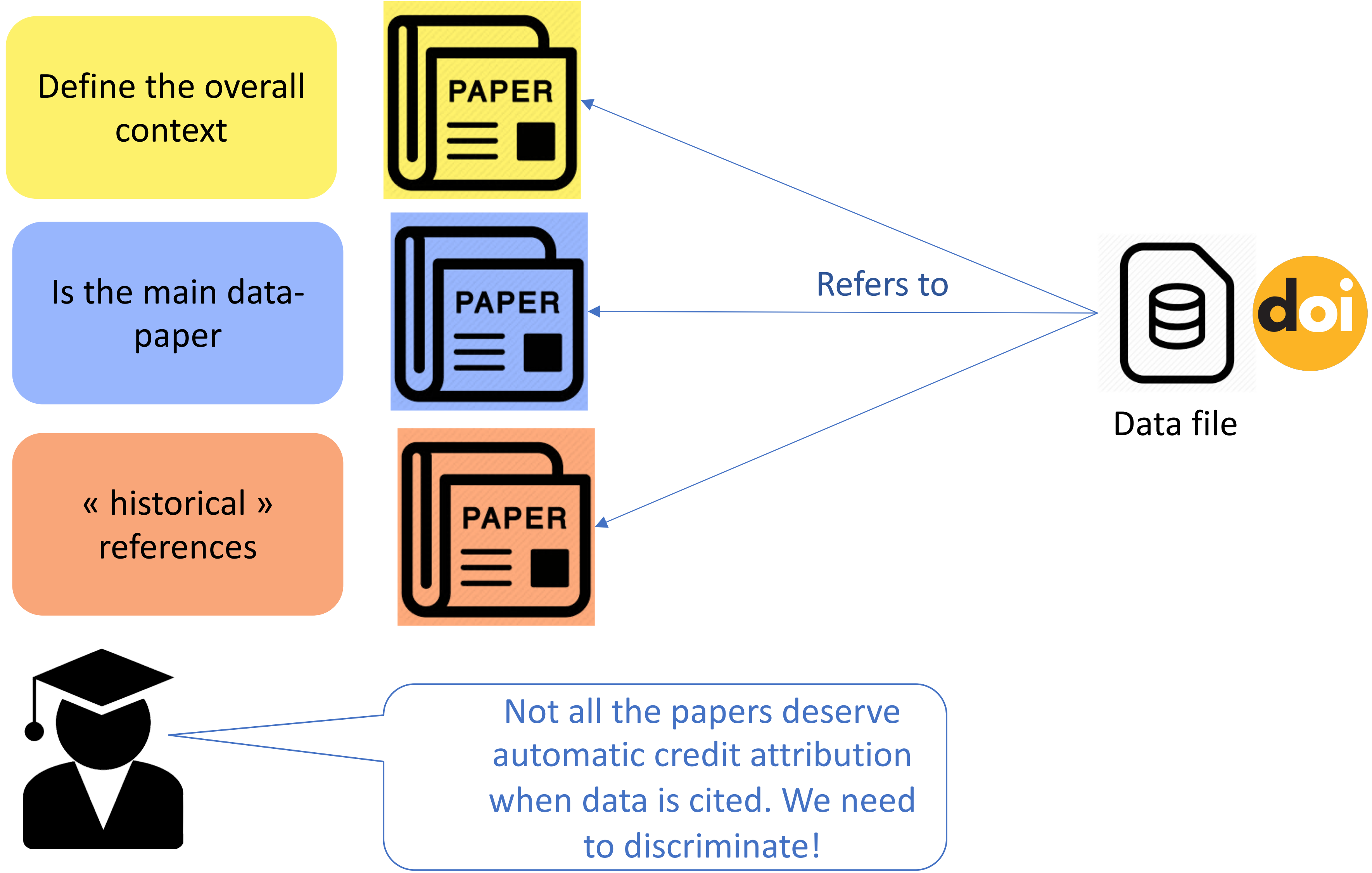
Define the overall context

Is the main data-paper

« historical » references

The referenced papers do not play the same role.

The sources of a new challenge



Define the overall context

Is the main data-paper

« historical » references

Refers to

Data file

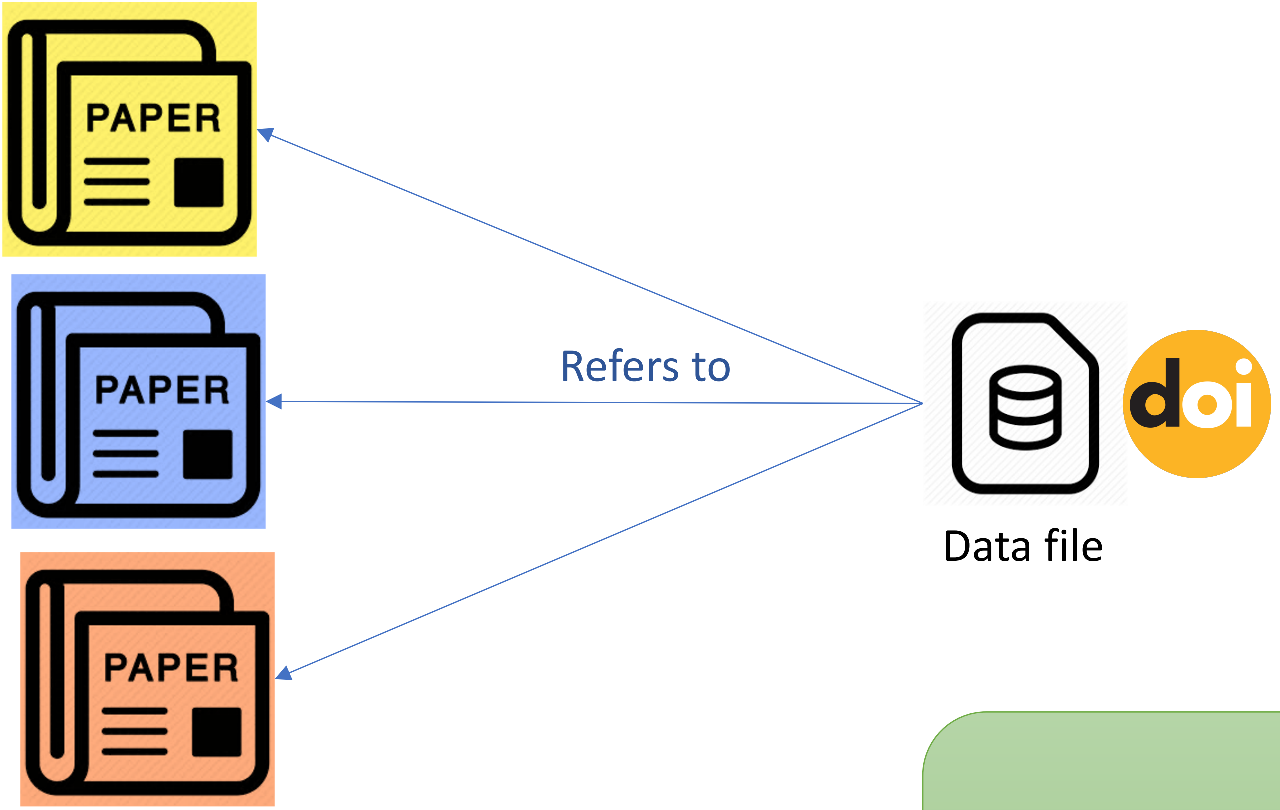
Not all the papers deserve automatic credit attribution when data is cited. We need to discriminate!

The sources of a new challenge

Define the overall context

Is the main data-paper

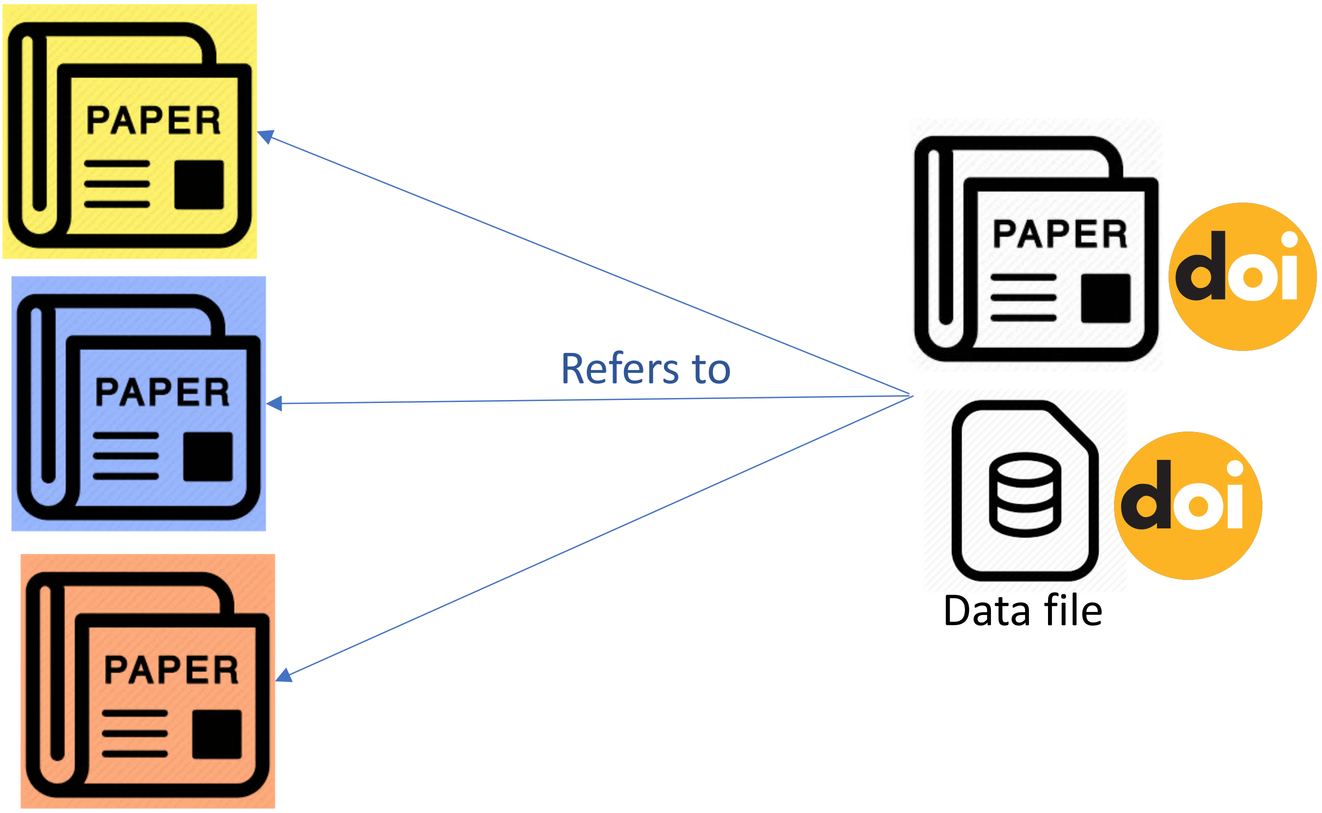
« historical » references



Not all the papers deserve automatic credit attribution when data is cited. We need to discriminate!

Need for rich metadata to describe the different citation contexts

The sources of a new challenge



Refers to

Data file

- This issue is common to both data-citation and « classic » paper citation.

- Consider a bibliography:
    - It contains no information about the citation context.
    - This can only been deduced from the text: only a human reader may understand it.

- Our aim is to provide the community with a mechanism for authors (both data and paper authors) to state the intent of a citation in a machine actionable way.

Capturing the intention behind a citation

Understanding the intention behind a citation is crucial for scientific reasons

Better attribution of bibliographic credits in automatic bibliometric workflows. But we are not interested in yet another H-factor-like indicator

- the reasons may provide a first assessment about the quality of what is cited.
  - A data-set which is cited as « crucial » in several other works presumably has a better quality compared to data-sets which has several citations from «erratum-works».
  - Let us consider for example the paper about the memory of water (doi: 10.1038/333816a0) which has a high H factor, but a lot of citations are (of course) negatives.

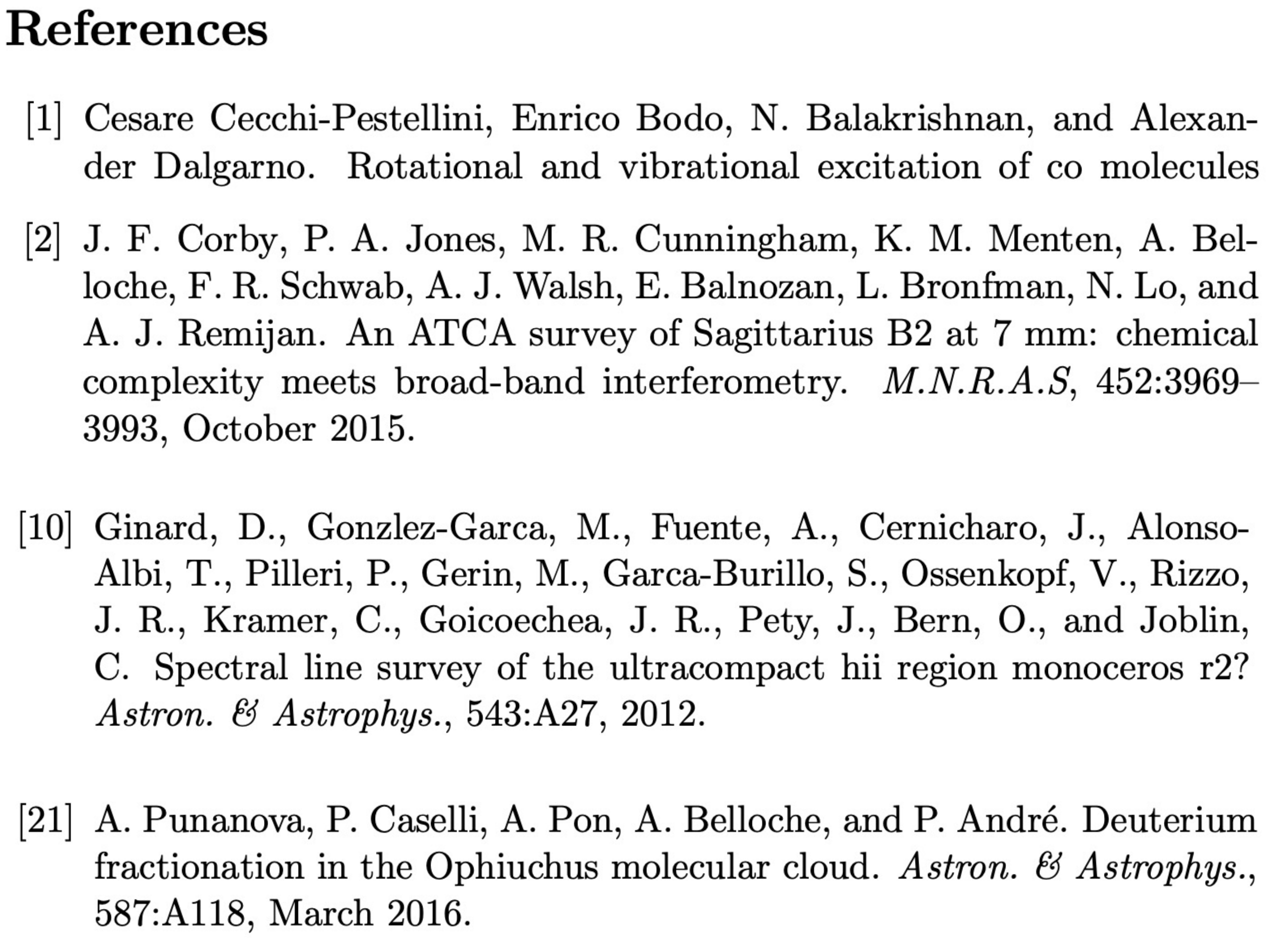
understanding how and why they work is re-used will help the data-producers to better fit the community needs.

Capturing the intention behind a citation

New model for datasets citation and extraction reproducibility in VAMDC
https://dx.doi.org/10.1016/j.jms.2016.04.009

Capturing the intention behind a citation

New model for datasets citation and extraction reproducibility in VAMDC
https://dx.doi.org/10.1016/j.jms.2016.04.009



References

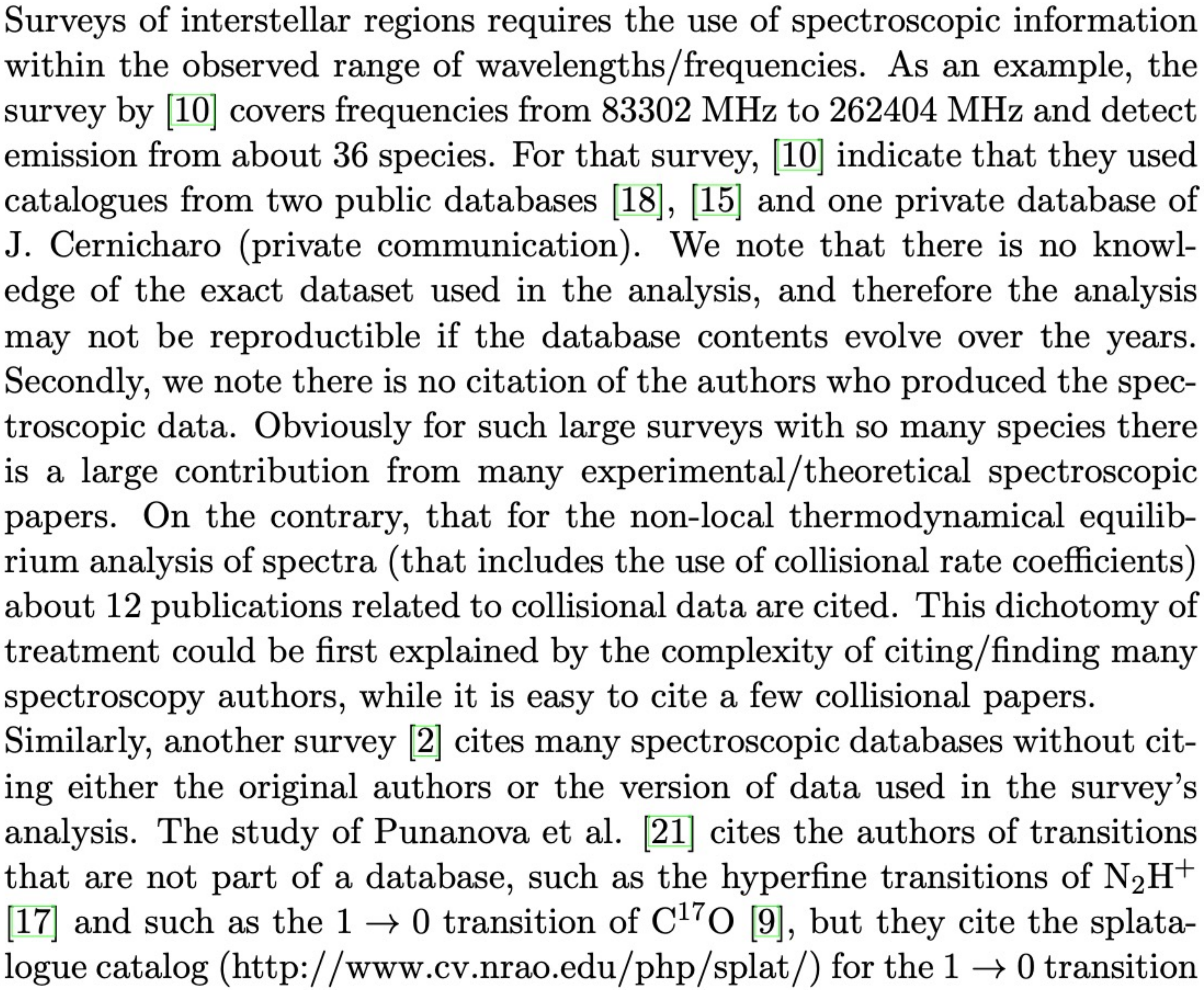
[1] Cesare Cecchi-Pestellini, Enrico Bodo, N. Balakrishnan, and Alexander Dalgarno. Rotational and vibrational excitation of co molecules

[2] J. F. Corby, P. A. Jones, M. R. Cunningham, K. M. Menten, A. Belloche, F. R. Schwab, A. J. Walsh, E. Balnozan, L. Bronfman, N. Lo, and A. J. Remijan. An ATCA survey of Sagittarius B2 at 7 mm: chemical complexity meets broad-band interferometry. *M.N.R.A.S*, 452:3969–3993, October 2015.

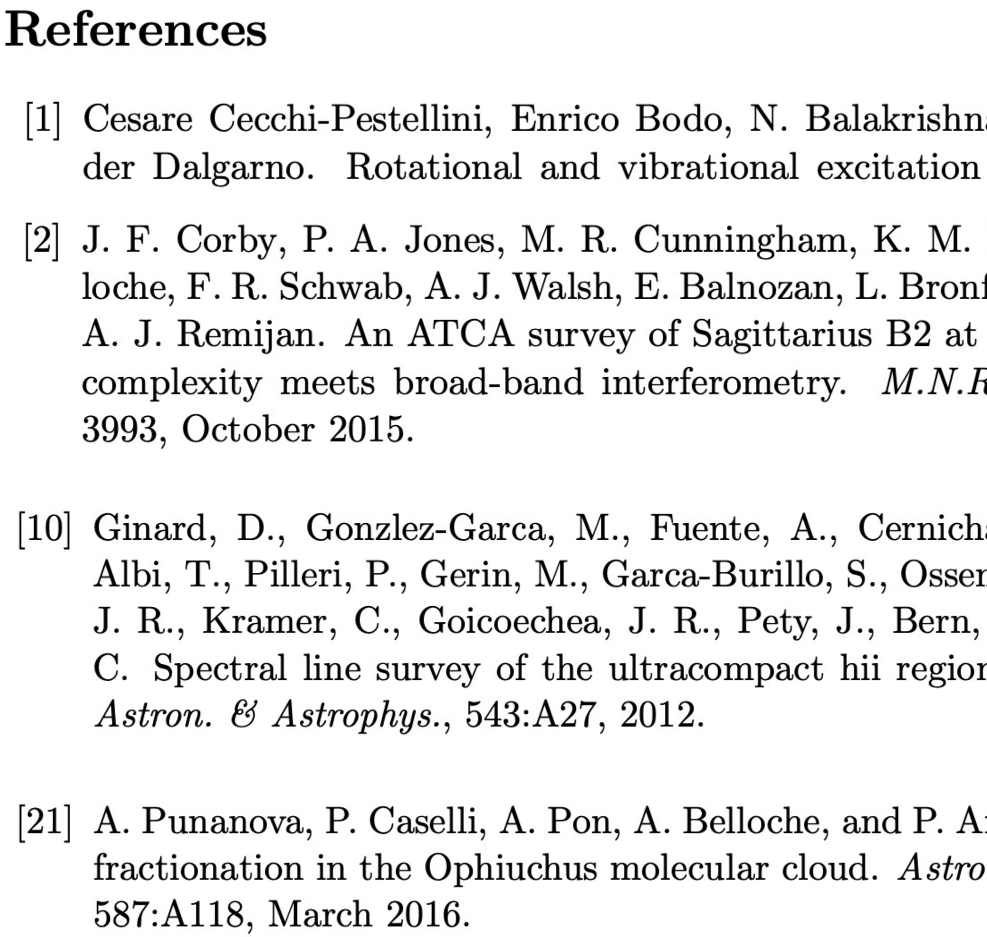
[10] Ginard, D., Gonzlez-Garca, M., Fuente, A., Cernicharo, J., Alonso-Albi, T., Pilleri, P., Gerin, M., Garca-Burillo, S., Ossenkopf, V., Rizzo, J. R., Kramer, C., Goicoechea, J. R., Pety, J., Bern, O., and Joblin, C. Spectral line survey of the ultracompact hii region monoceros r2? *Astron. & Astrophys.*, 543:A27, 2012.

[21] A. Punanova, P. Caselli, A. Pon, A. Belloche, and P. André. Deuterium fractionation in the Ophiuchus molecular cloud. *Astron. & Astrophys.*, 587:A118, March 2016.

Capturing the intention behind a citation



Surveys of interstellar regions requires the use of spectroscopic information within the observed range of wavelengths/frequencies. As an example, the survey by [10] covers frequencies from 83302 MHz to 262404 MHz and detect emission from about 36 species. For that survey, [10] indicate that they used catalogues from two public databases [18], [15] and one private database of J. Cernicharo (private communication). We note that there is no knowledge of the exact dataset used in the analysis, and therefore the analysis may not be reproductible if the database contents evolve over the years. Secondly, we note there is no citation of the authors who produced the spectroscopic data. Obviously for such large surveys with so many species there is a large contribution from many experimental/theoretical spectroscopic papers. On the contrary, that for the non-local thermodynamical equilibrium analysis of spectra (that includes the use of collisional rate coefficients) about 12 publications related to collisional data are cited. This dichotomy of treatment could be first explained by the complexity of citing/finding many spectroscopy authors, while it is easy to cite a few collisional papers.
Similarly, another survey [2] cites many spectroscopic databases without citing either the original authors or the version of data used in the survey's analysis. The study of Punanova et al. [21] cites the authors of transitions that are not part of a database, such as the hyperfine transitions of N_2H^+ [17] and such as the 1 → 0 transition of $C^{17}O$ [9], but they cite the splatalogue catalog (http://www.cv.nrao.edu/php/splat/) for the 1 → 0 transition



References

[1] Cesare Cecchi-Pestellini, Enrico Bodo, N. Balakrishn
der Dalgarno. Rotational and vibrational excitation

[2] J. F. Corby, P. A. Jones, M. R. Cunningham, K. M.
loche, F. R. Schwab, A. J. Walsh, E. Balnozan, L. Bron
A. J. Remijan. An ATCA survey of Sagittarius B2 at
complexity meets broad-band interferometry. *M.N.R*
3993, October 2015.

[10] Ginard, D., Gonzlez-Garca, M., Fuente, A., Cernich
Albi, T., Pilleri, P., Gerin, M., Garca-Burillo, S., Osse
J. R., Kramer, C., Goicoechea, J. R., Pety, J., Bern,
C. Spectral line survey of the ultracompact hii regio
Astron. & Astrophys., 543:A27, 2012.

[21] A. Punanova, P. Caselli, A. Pon, A. Belloche, and P. A
fractionation in the Ophiuchus molecular cloud. *Astro*
587:A118, March 2016.

Capturing the intention behind a citation

Existing solutions?

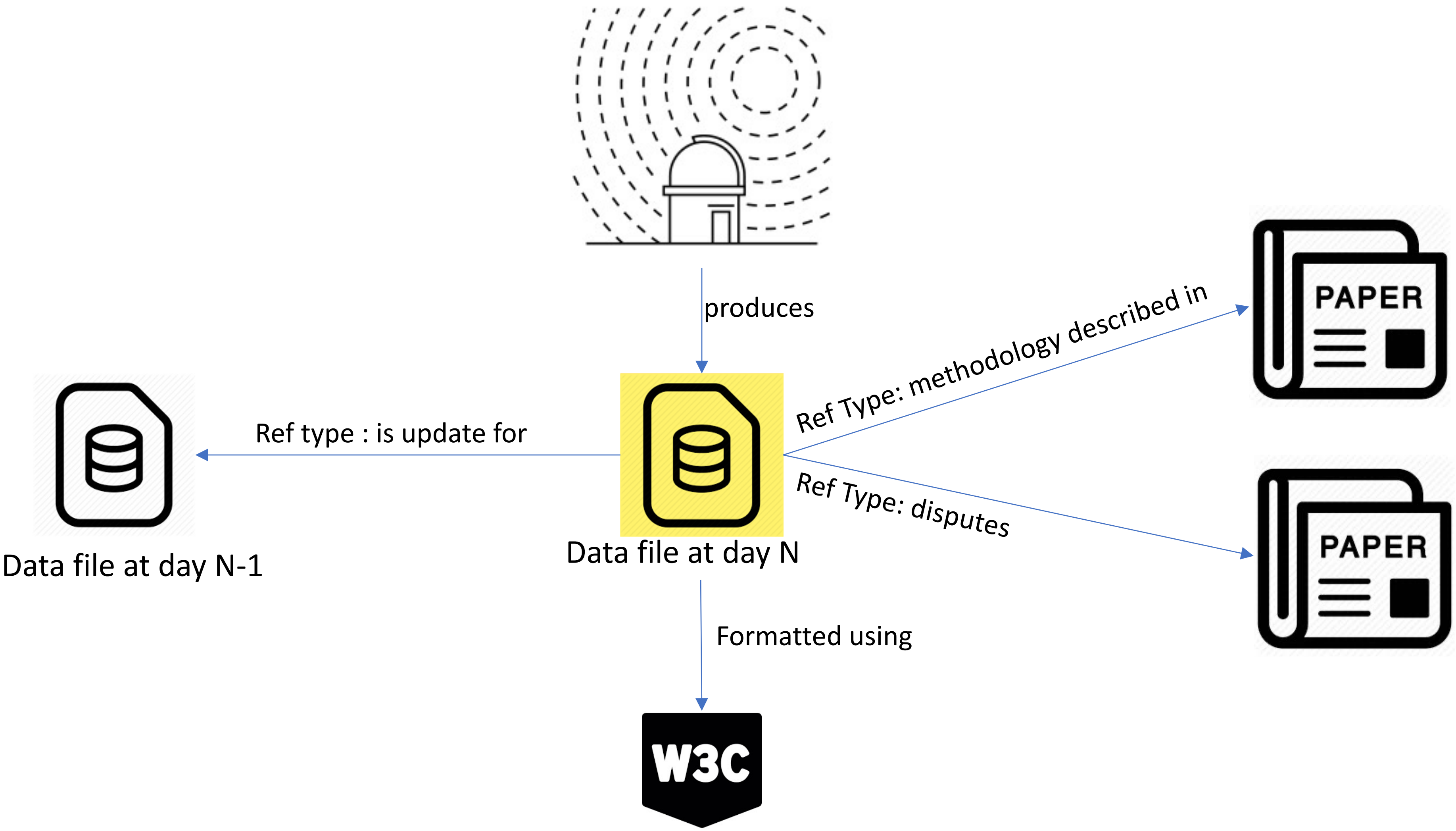
No existing standard to annotate citation context/intention in a machine actionable way

- Interesting ideas are published in literature:
  - Before 2000: Bibliometrics papers
  - After 2000: *Natural Language processing* & *machine learning* for classifying citations. The definition of categories is part of these works
- Nobody succeeded in creating a momentum around a particular solution

- We proposed a BoF at the next RDA plenary to create this momentum: https://www.rd-alliance.org/rich-metadata-annotation-citations-contexts-and-data-citations-contexts

- Our aim is to provide the community with a mechanism for authors (both data and paper authors) to state the intent of a citation in a machine actionable way.

Some examples



produces

Ref type : is update for

Ref Type: methodology described in

Ref Type: disputes

Data file at day N-1

Data file at day N

Formatted using

Some examples

Contents lists available at SciVerse ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

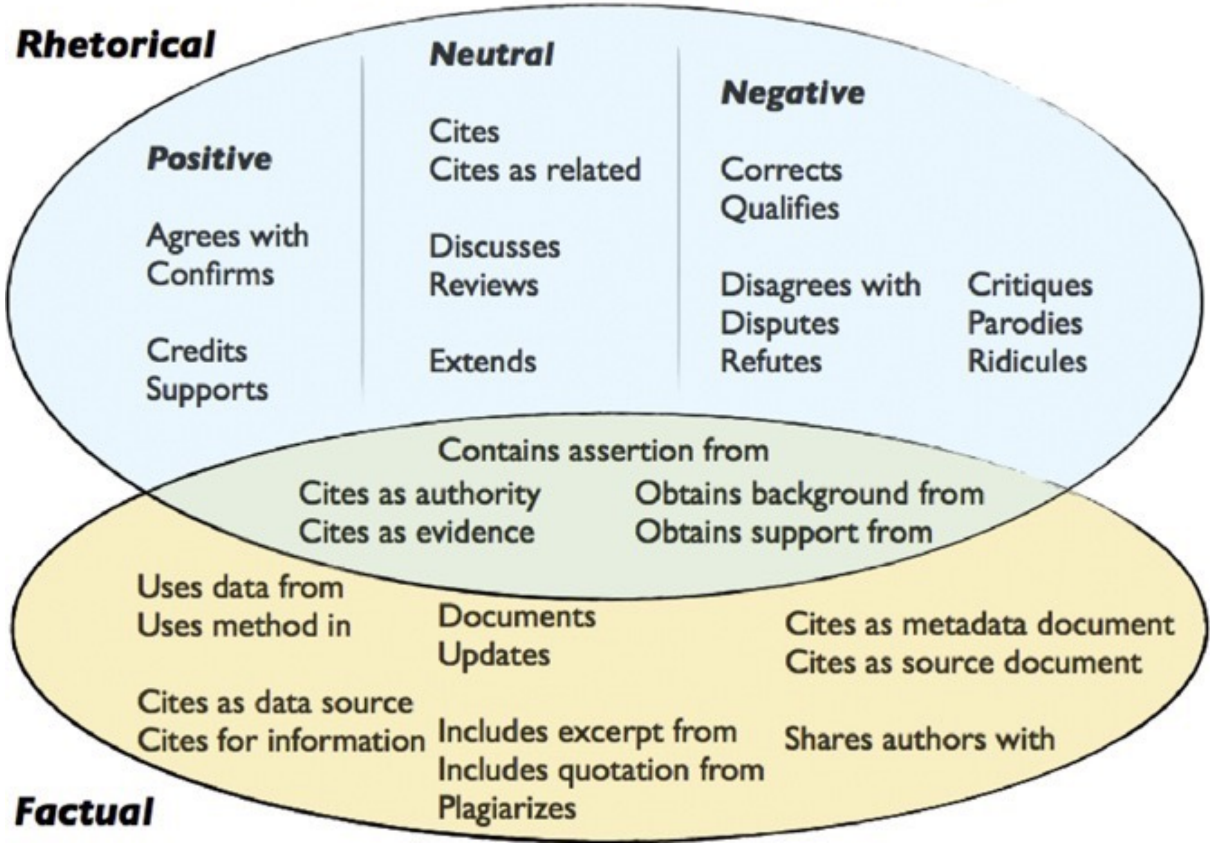
journal homepage: www.elsevier.com/locate/websem

Ontology paper

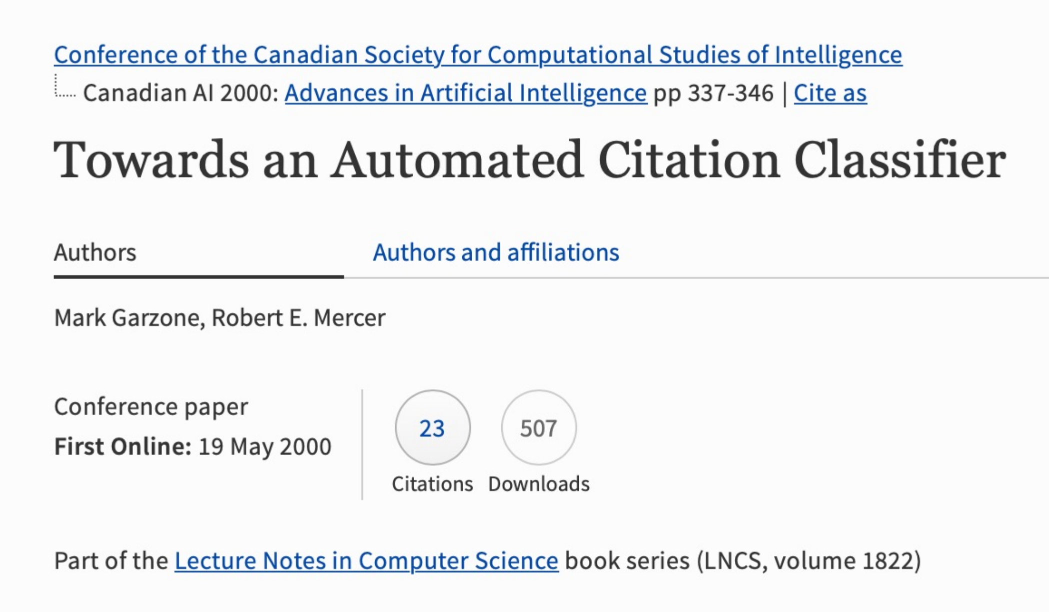
FaBiO and CiTO: Ontologies for describing bibliographic resources and citations

Silvio Peroni[a,*], David Shotton[b]

Clustering of CiTO relationships by similarity



Some examples



Conference of the Canadian Society for Computational Studies of Intelligence
Canadian AI 2000: Advances in Artificial Intelligence pp 337-346 | Cite as

Towards an Automated Citation Classifier

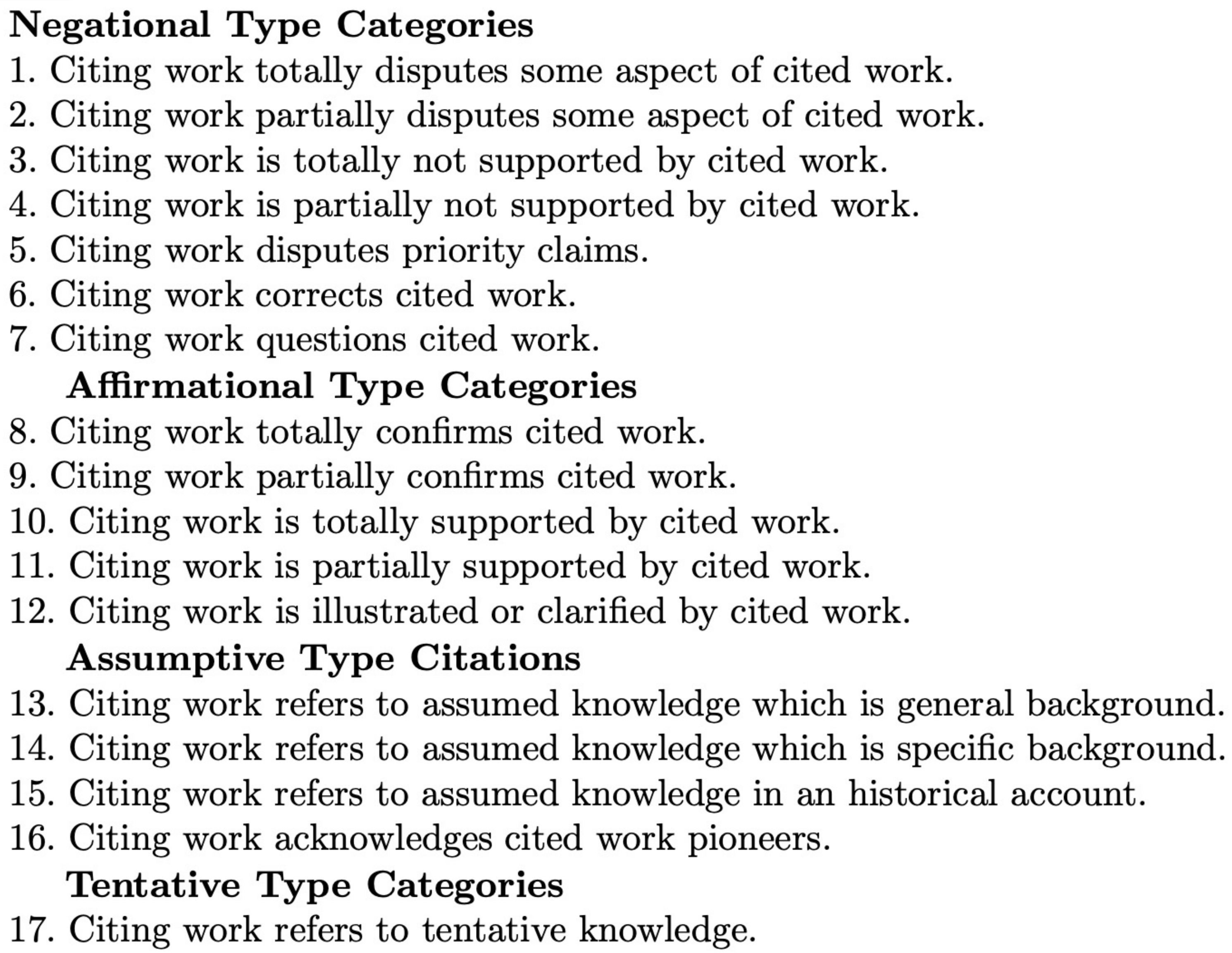
Authors | Authors and affiliations

Mark Garzone, Robert E. Mercer

Conference paper
First Online: 19 May 2000

23 Citations | 507 Downloads

Part of the Lecture Notes in Computer Science book series (LNCS, volume 1822)



Negational Type Categories

1. Citing work totally disputes some aspect of cited work.
2. Citing work partially disputes some aspect of cited work.
3. Citing work is totally not supported by cited work.
4. Citing work is partially not supported by cited work.
5. Citing work disputes priority claims.
6. Citing work corrects cited work.
7. Citing work questions cited work.

Affirmational Type Categories

8. Citing work totally confirms cited work.
9. Citing work partially confirms cited work.
10. Citing work is totally supported by cited work.
11. Citing work is partially supported by cited work.
12. Citing work is illustrated or clarified by cited work.

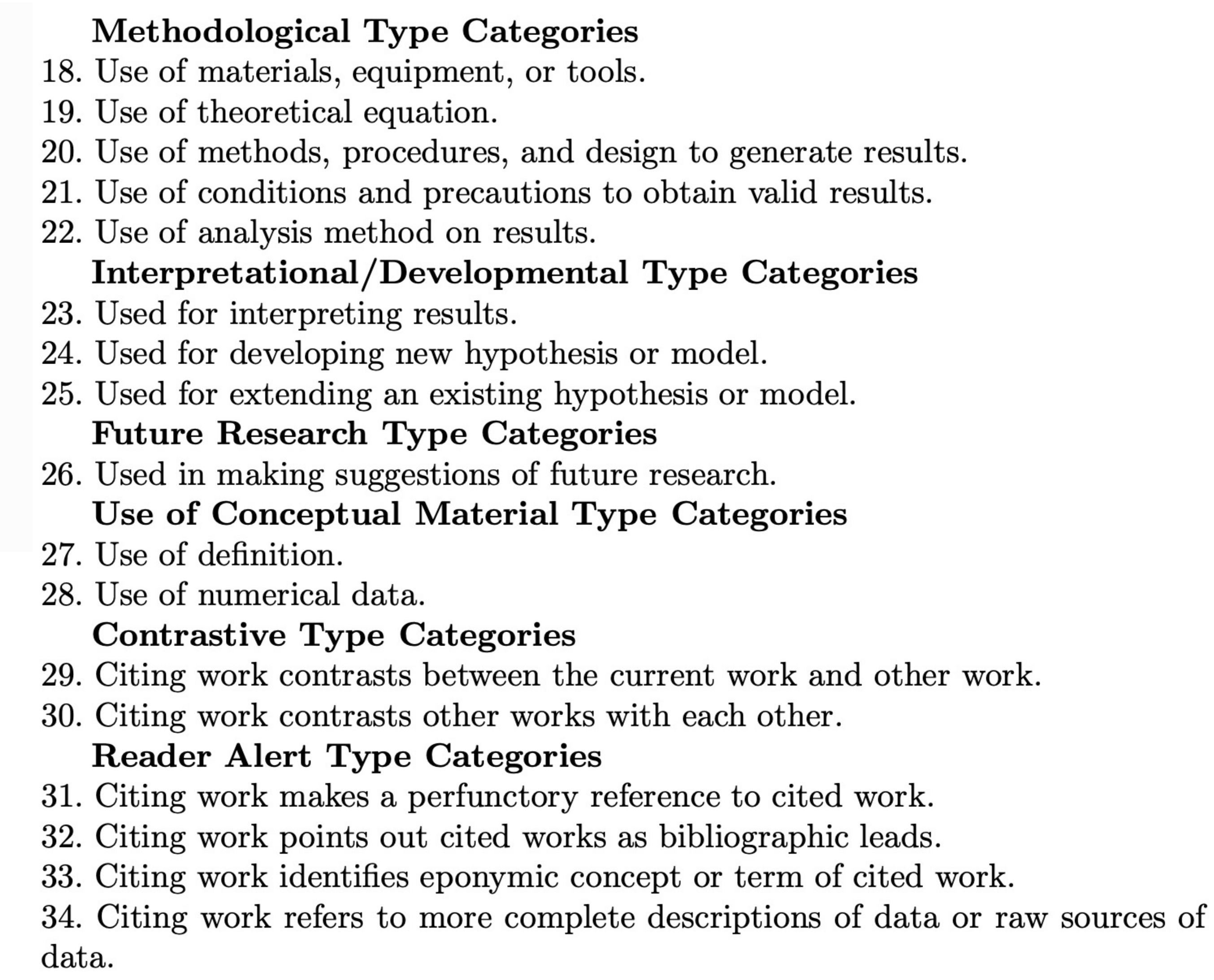
Assumptive Type Citations

13. Citing work refers to assumed knowledge which is general background.
14. Citing work refers to assumed knowledge which is specific background.
15. Citing work refers to assumed knowledge in an historical account.
16. Citing work acknowledges cited work pioneers.

Tentative Type Categories

17. Citing work refers to tentative knowledge.

Some examples



Methodological Type Categories

18. Use of materials, equipment, or tools.
19. Use of theoretical equation.
20. Use of methods, procedures, and design to generate results.
21. Use of conditions and precautions to obtain valid results.
22. Use of analysis method on results.

Interpretational/Developmental Type Categories

23. Used for interpreting results.
24. Used for developing new hypothesis or model.
25. Used for extending an existing hypothesis or model.

Future Research Type Categories

26. Used in making suggestions of future research.

Use of Conceptual Material Type Categories

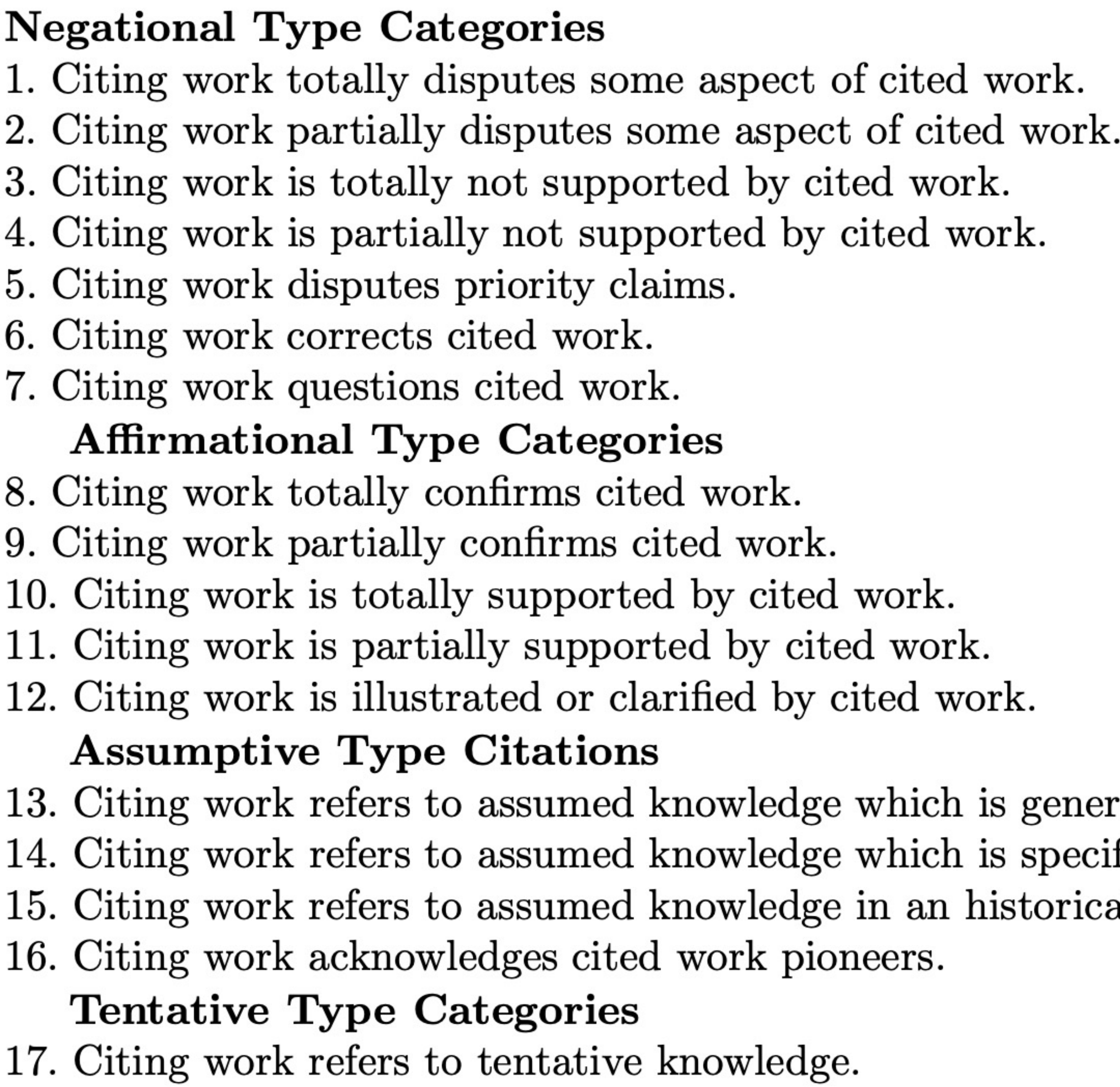
27. Use of definition.
28. Use of numerical data.

Contrastive Type Categories

29. Citing work contrasts between the current work and other work.
30. Citing work contrasts other works with each other.

Reader Alert Type Categories

31. Citing work makes a perfunctory reference to cited work.
32. Citing work points out cited works as bibliographic leads.
33. Citing work identifies eponymic concept or term of cited work.
34. Citing work refers to more complete descriptions of data or raw sources of data.



Negational Type Categories

1. Citing work totally disputes some aspect of cited work.
2. Citing work partially disputes some aspect of cited work.
3. Citing work is totally not supported by cited work.
4. Citing work is partially not supported by cited work.
5. Citing work disputes priority claims.
6. Citing work corrects cited work.
7. Citing work questions cited work.

Affirmational Type Categories

8. Citing work totally confirms cited work.
9. Citing work partially confirms cited work.
10. Citing work is totally supported by cited work.
11. Citing work is partially supported by cited work.
12. Citing work is illustrated or clarified by cited work.

Assumptive Type Citations

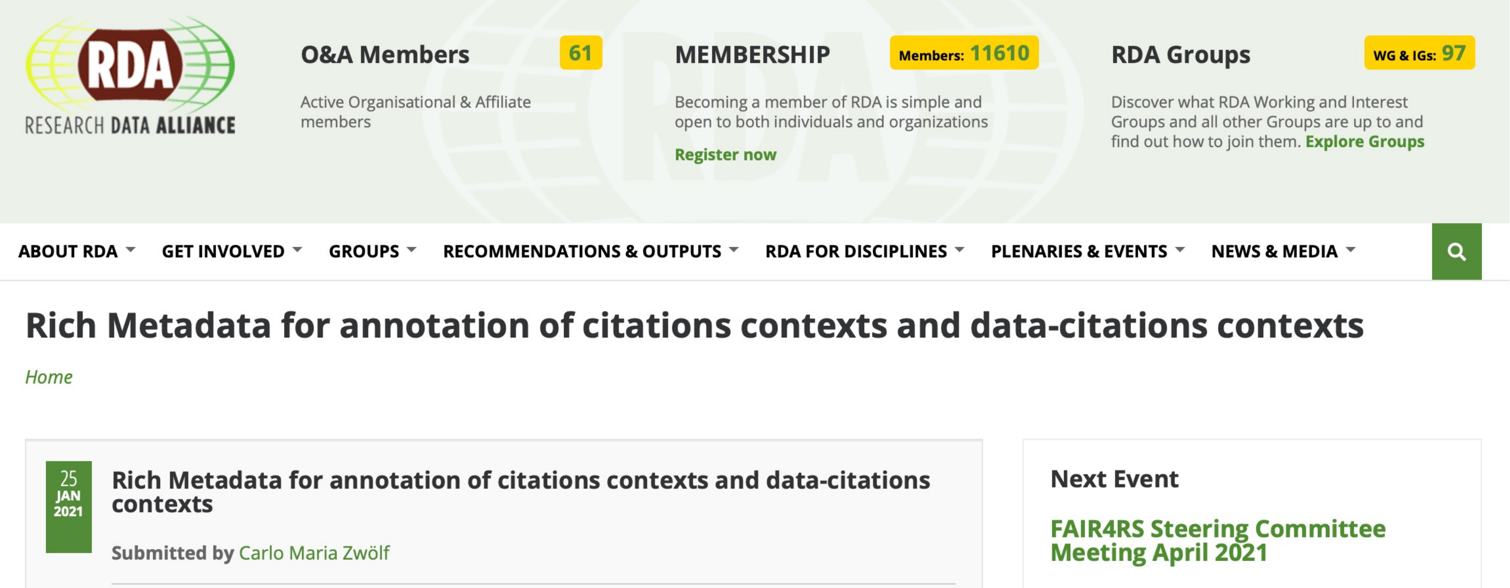
13. Citing work refers to assumed knowledge which is gener
14. Citing work refers to assumed knowledge which is specif
15. Citing work refers to assumed knowledge in an historica
16. Citing work acknowledges cited work pioneers.

Tentative Type Categories

17. Citing work refers to tentative knowledge.

Conclusions

- At the next RDA plenary we will try to establish a new working group for addressing these issues
- Create a momentum to aggregate the existing scattered solution fragments into a community standard



Active Organisational & Affiliate members

Becoming a member of RDA is simple and open to both individuals and organizations

Discover what RDA Working and Interest Groups and all other Groups are up to and find out how to join them.

Rich Metadata for annotation of citations contexts and data-citations contexts

Rich Metadata for annotation of citations contexts and data-citations contexts

Submitted by Carlo Maria Zwölf

FAIR4RS Steering Committee Meeting April 2021

See you soon at the RDA plenary…