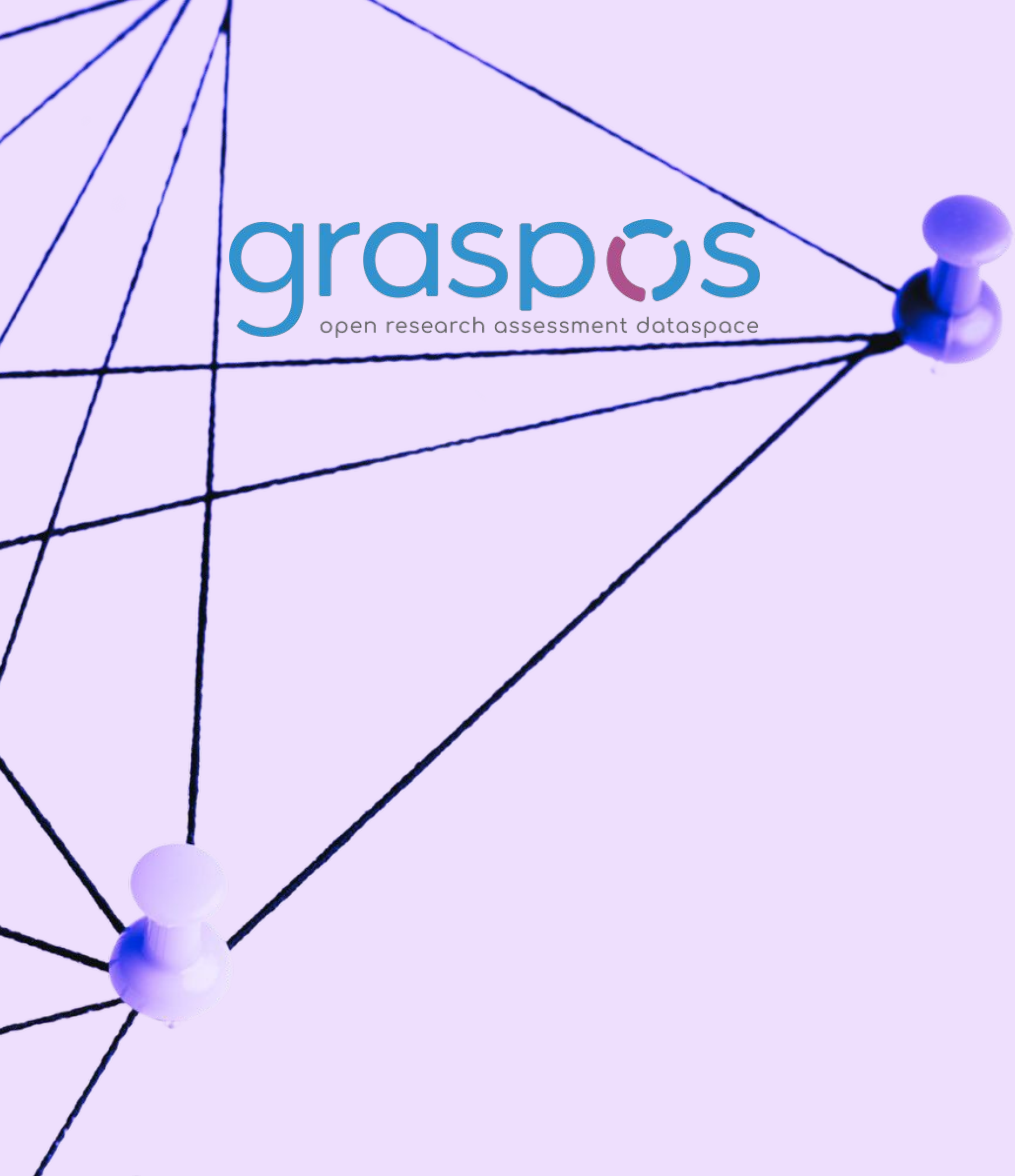


graspos
open research assessment dataspace

Citations and their meaning

– or why we cite

Silvio Peroni

Research Centre for Open Scholarly Metadata,
Department of Classical Philology and Italian Studies, University of Bologna, Italy
silvio.peroni@unibo.it – @essepuntato@scholar.social – https://www.unibo.it/sitoweb/silvio.peroni/en

OpenCitations, Director
silvio.peroni@opencitations.net – @opencitations@scicomm.xyz – https://opencitations.net

RESEARCH CENTRE
FOR OPEN SCHOLARLY METADATA

OpenCitations

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

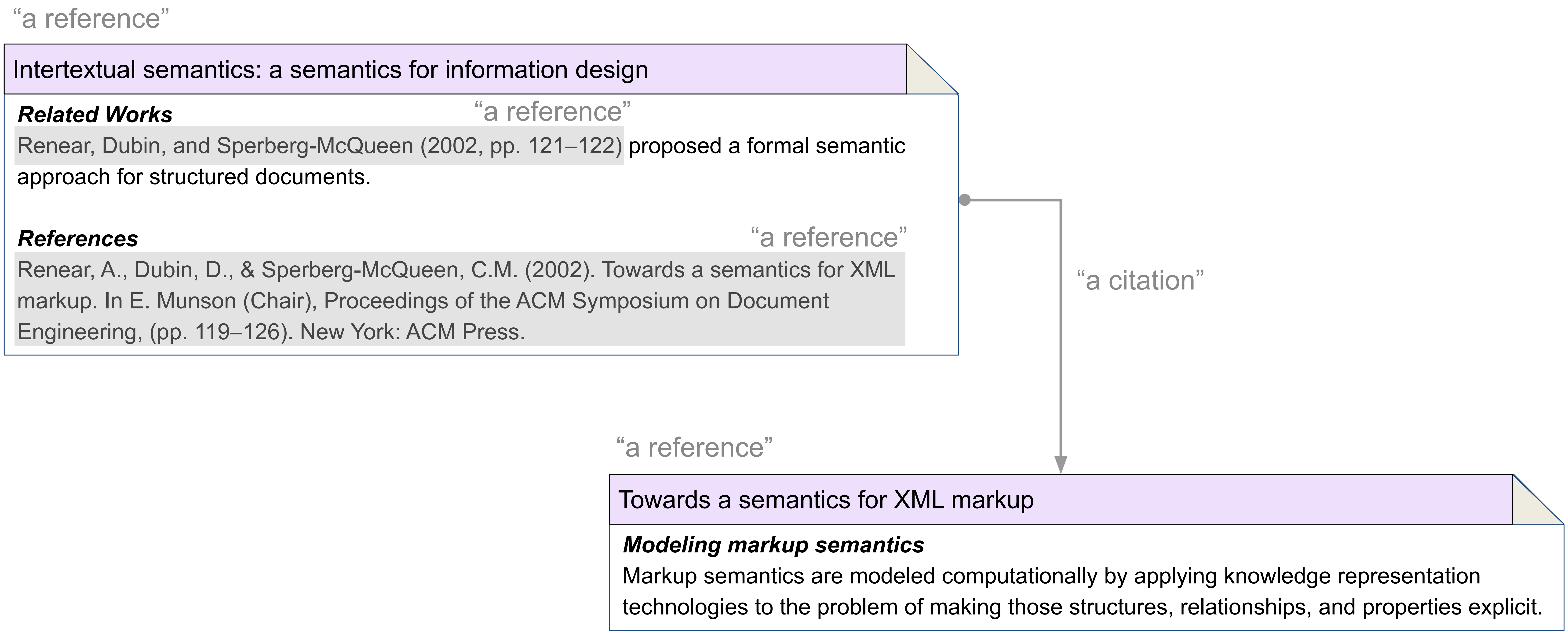
Acknowledgments

This work could not be possible without the enthusiastic and professional collaboration of

- Angelo Di Iorio
- Ivan Heibi
- Olga Pagnotta
- Lorenzo Paolini
- Marta Soricetti

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

“A reference”



“a reference”

Intertextual semantics: a semantics for information design

Related Works

“a reference”

Renear, Dubin, and Sperberg-McQueen (2002, pp. 121–122) proposed a formal semantic approach for structured documents.

References

“a reference”

Renear, A., Dubin, D., & Sperberg-McQueen, C.M. (2002). Towards a semantics for XML markup. In E. Munson (Chair), Proceedings of the ACM Symposium on Document Engineering, (pp. 119–126). New York: ACM Press.

“a citation”

“a reference”

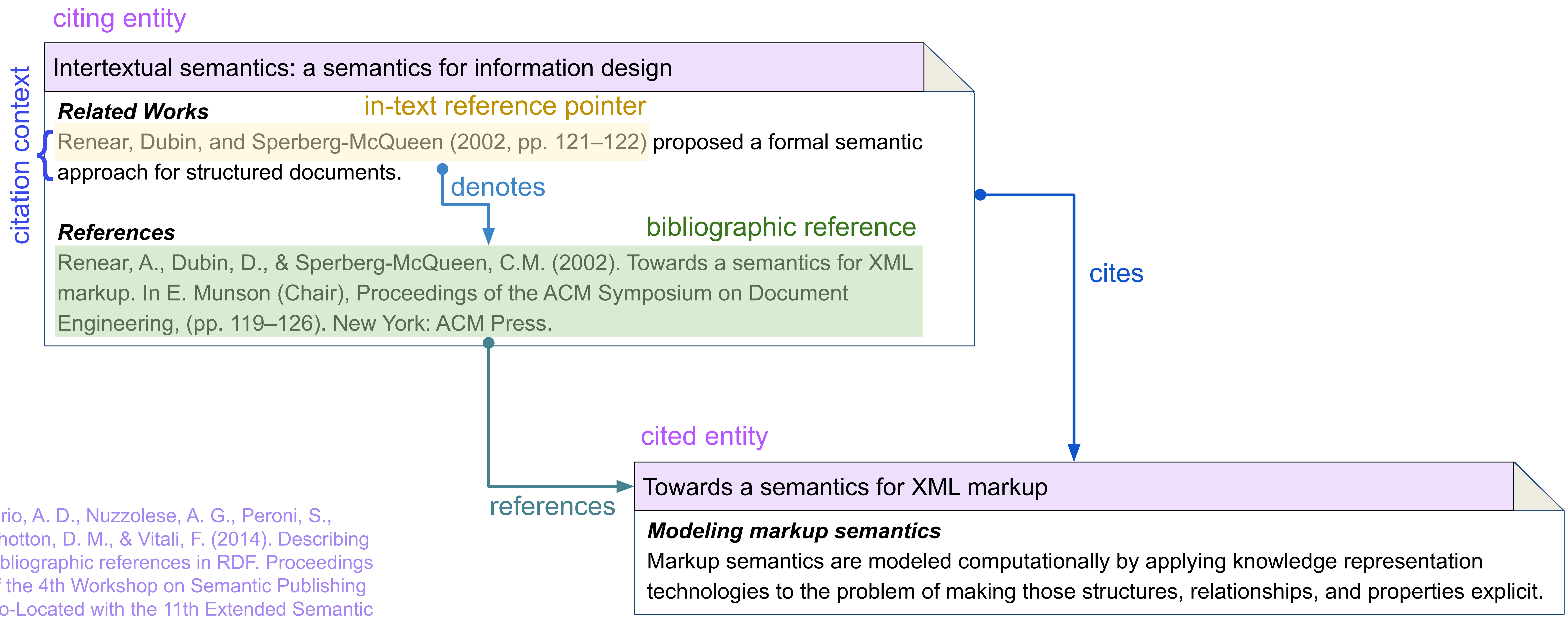
Towards a semantics for XML markup

Modeling markup semantics

Markup semantics are modeled computationally by applying knowledge representation technologies to the problem of making those structures, relationships, and properties explicit.

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

We need to use appropriate terms



citing entity

citation context

Intertextual semantics: a semantics for information design

Related Works

in-text reference pointer

Renear, Dubin, and Sperberg-McQueen (2002, pp. 121–122) proposed a formal semantic approach for structured documents.

denotes

References

bibliographic reference

Renear, A., Dubin, D., & Sperberg-McQueen, C.M. (2002). Towards a semantics for XML markup. In E. Munson (Chair), Proceedings of the ACM Symposium on Document Engineering, (pp. 119–126). New York: ACM Press.

cites

references

cited entity

Towards a semantics for XML markup

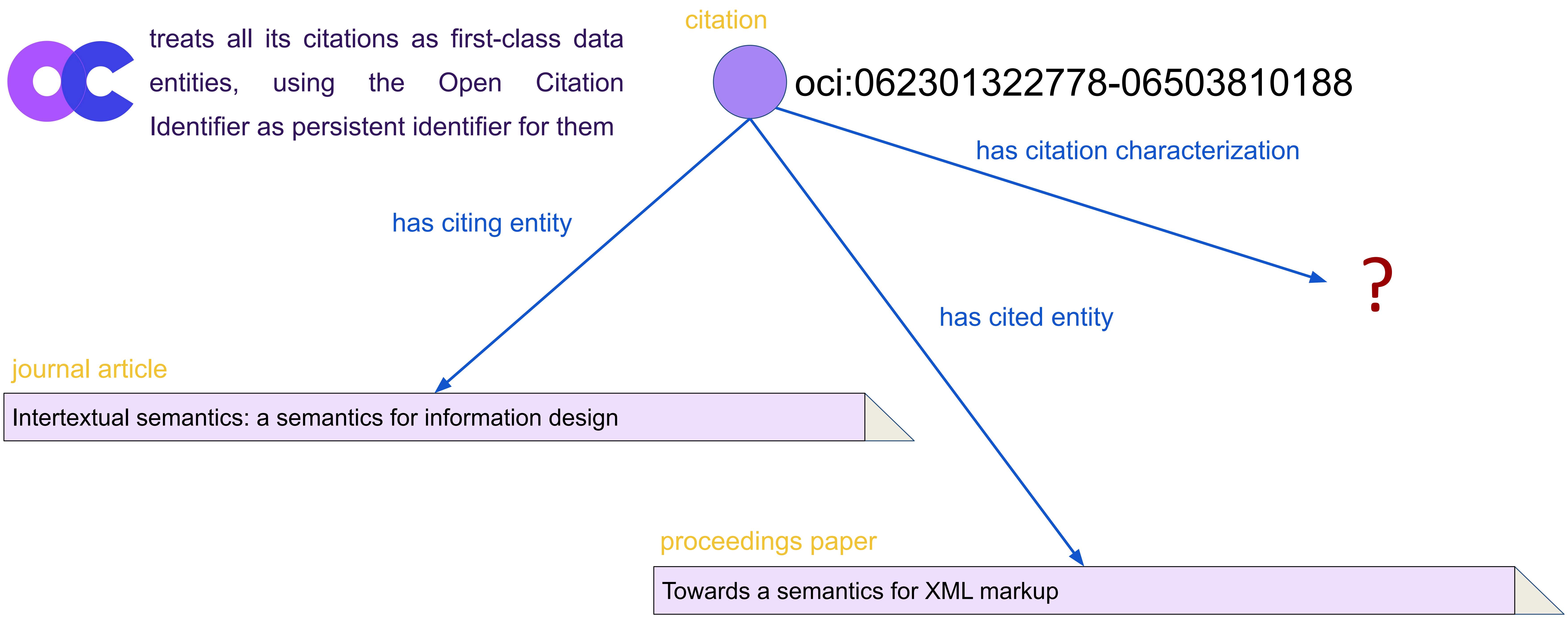
Modeling markup semantics

Markup semantics are modeled computationally by applying knowledge representation technologies to the problem of making those structures, relationships, and properties explicit.

Iorio, A. D., Nuzzolese, A. G., Peroni, S., Shotton, D. M., & Vitali, F. (2014). Describing bibliographic references in RDF. Proceedings of the 4th Workshop on Semantic Publishing Co-Located with the 11th Extended Semantic Web Conference (ESWC 2014), Anissaras, Greece, May 25th, 2014., 1155. http://ceur-ws.org/Vol-1155/paper-05.pdf

Citations as first-class data entities

treats all its citations as first-class data entities, using the Open Citation Identifier as persistent identifier for them


citation
oci:062301322778-06503810188
has citing entity
has cited entity
has citation characterization
?
journal article
Intertextual semantics: a semantics for information design
proceedings paper
Towards a semantics for XML markup

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

We cite for a reason

“Citation function is defined as the author’s reason for citing a given paper (e.g. acknowledgement of the use of the cited method)”

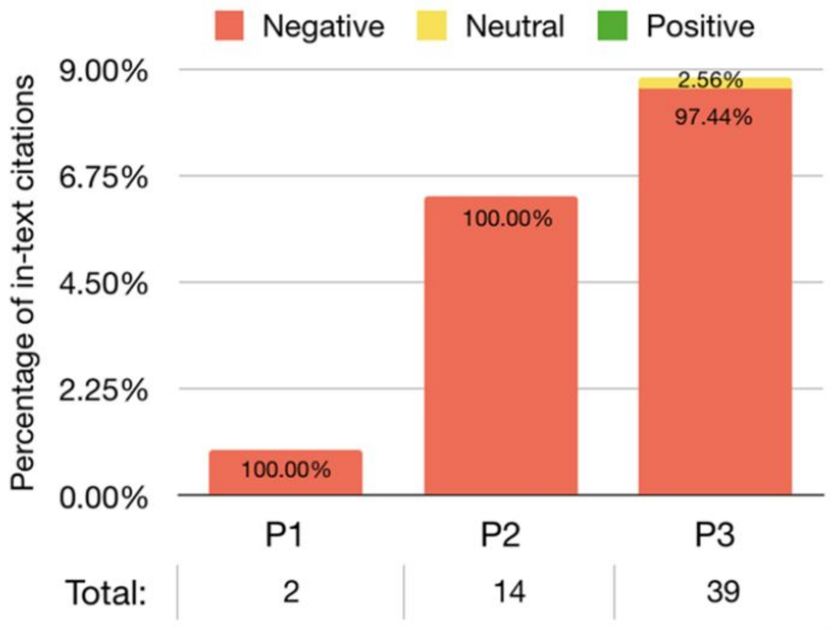
Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP ’06), 103. https://doi.org/10.3115/1610075.1610091

In the past, plenty of different citation annotation schemes has been proposed, which have been developed for catching different dimensions

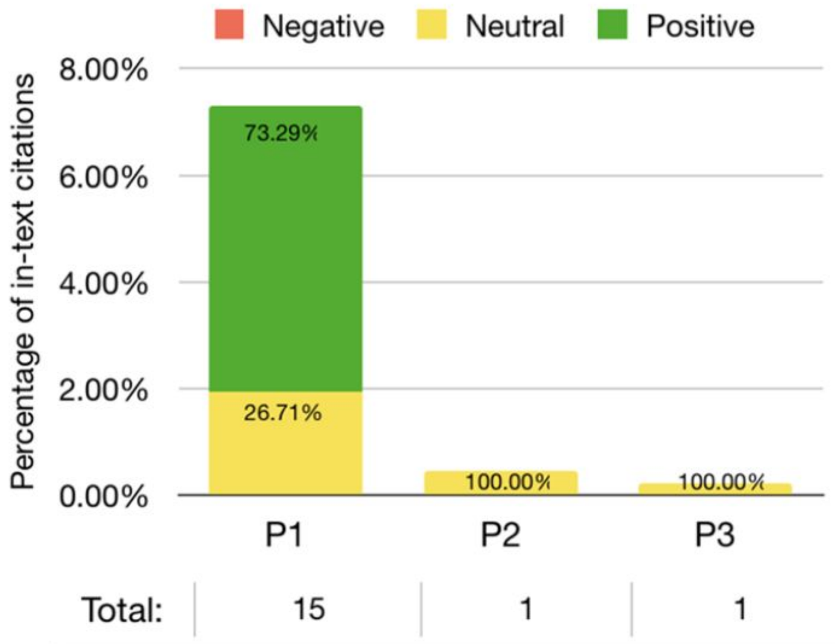
Kunnath, S. N., Herrmannova, D., Pride, D., & Knoth, P. (2021). A meta-analysis of semantic classification of citations. Quantitative Science Studies, 2(4), 1170–1215. https://doi.org/10.1162/qss_a_00159

Having such a *labelled* graph of citations enables us to study how different articles and/or authors interact with each other and identify patterns on how one work is relevant for the community

Heibi, I., & Peroni, S. (2021). A qualitative and quantitative analysis of open citations to retracted articles: The Wakefield et al.’s case. Scientometrics, 126(10), 8433–8470. https://doi.org/10.1007/s11192-021-04097-5



critiques



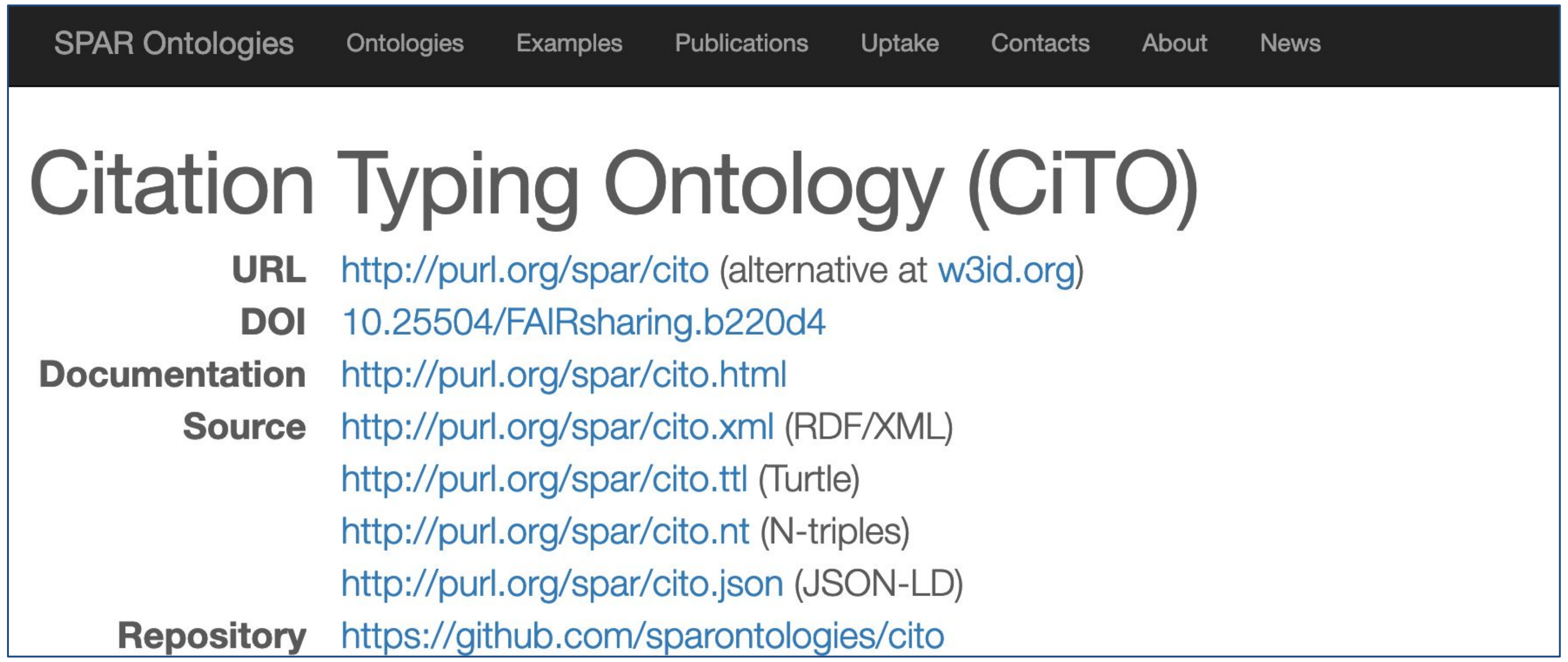
obtains support from + uses conclusions from + extends + updates + uses data from

An ontology

CiTO, the Citation Typing Ontology, makes it possible to mark citation links and to capture their citation intent (e.g. extends, uses method in, supports) when someone cites a particular entity – more than 40 intents available!

CiTO allows one to create metadata describing citations that are distinct from metadata describing the cited works themselves, and permits the motives of an author when referring to another document to be captured

http://www.sparontologies.net



Citation Typing Ontology (CiTO)

URL	http://purl.org/spar/cito (alternative at w3id.org)
DOI	10.25504/FAIRsharing.b220d4
Documentation	http://purl.org/spar/cito.html
Source	http://purl.org/spar/cito.xml (RDF/XML) http://purl.org/spar/cito.ttl (Turtle) http://purl.org/spar/cito.nt (N-triples) http://purl.org/spar/cito.json (JSON-LD)
Repository	https://github.com/sparontologies/cito

An example in RDF (Turtle) format

```
@prefix cito: <https://purl.org/spar/cito/> .
@prefix oci: <https://w3id.org/oc/index/ci/> .
@prefix omid: <https://w3id.org/oc/meta/br/> .

oci:062301322778-06503810188 a cito:Citation ;
    cito:hasCitingEntity omid:062301322778 ;
    cito:hasCitedEntity omid:06503810188 ;
    cito:hasCitationCharacterization cito:citesAsRelated .
```

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Journal of Web Semantics, 17, 33–43. https://doi.org/10.1016/j.websem.2012.08.001

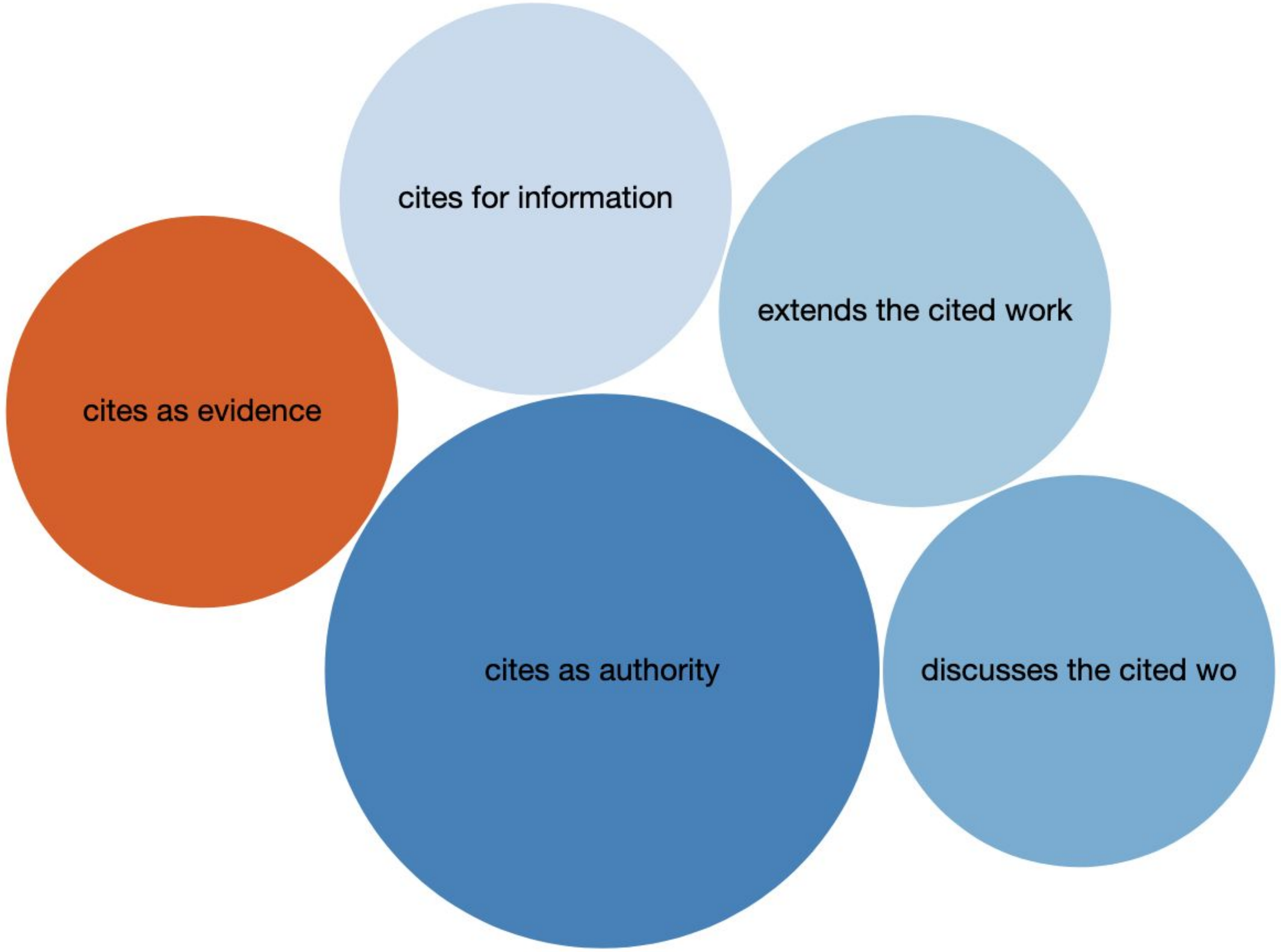
Adoption of CiTO

In the past years, we have seen a few experimentation on using CiTO into specific publication-oriented scenario

The Journal of Cheminformatics run a pilot where authors where allowed to accompanied the reference lists of their articles with the reason for citing via in-text annotations

Willighagen, E. (2023). Two years of explicit CiTO annotations. Journal of Cheminformatics, 15(1), 14. https://doi.org/10.1186/s13321-023-00683-2

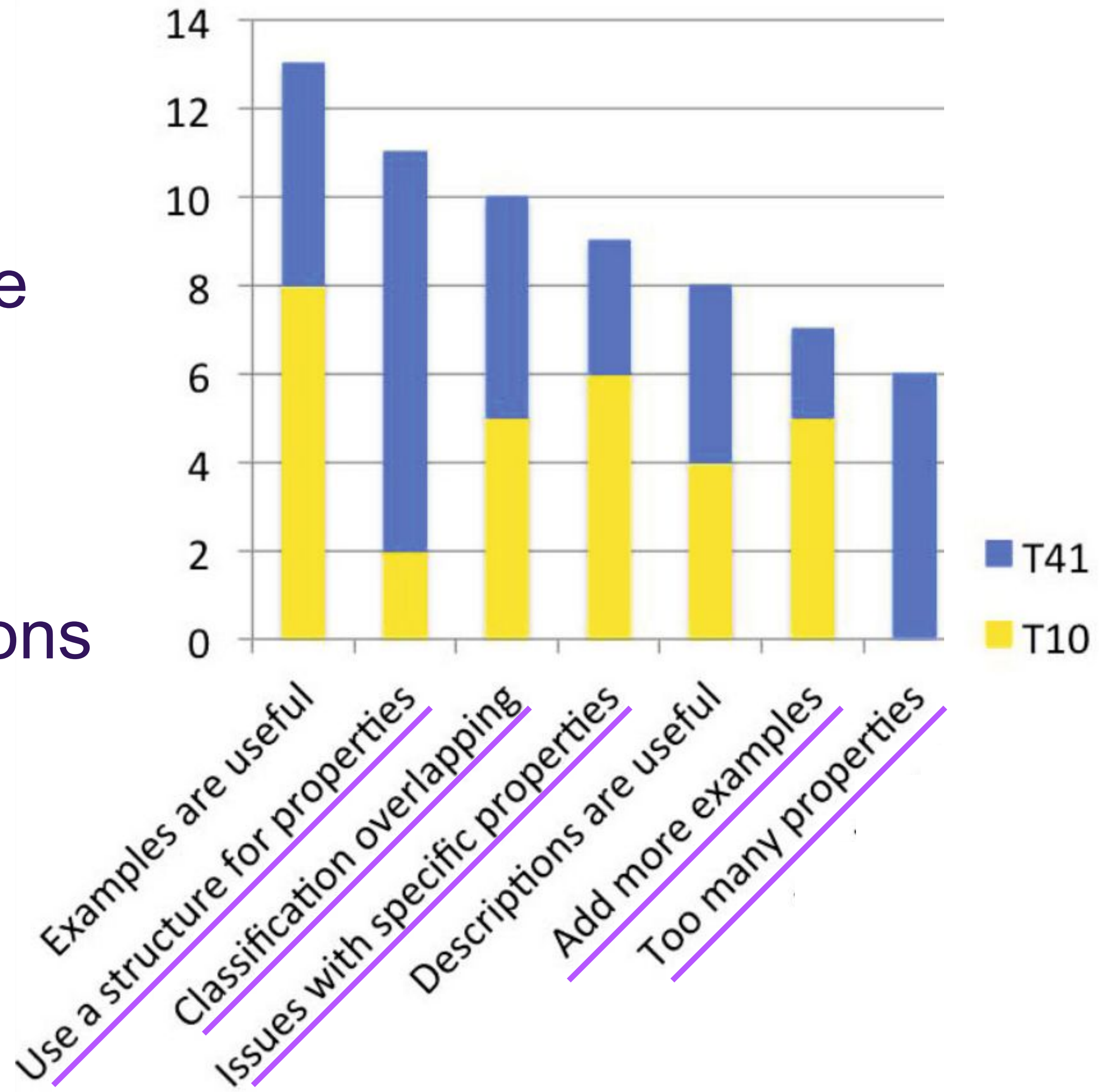
Scholia (https://scholia.toolforge.org/) used the citation intents annotated in Wikidata (via CiTO) and expose them to its interface when available



Problems in annotating citation intents

To create large dataset of annotations of citation intents, we need annotations done by either authors or readers, but manual annotation is very time consuming and does not scale well

In addition, we have measured how human readers annotate scholarly article with CiTO, by measuring the annotation agreement in 105 citations performed by 20 people, with 10 annotators for each citations and within two different conditions (i.e. use either 41 or 10 CiTO citation intents)



The inter-rater agreement was low for both conditions

Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2014). Evaluating Citation Functions in CiTO: Cognitive Issues. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, & A. Tordai (Eds.), The Semantic Web: Trends and Challenges (Vol. 8465, pp. 580–594). Springer International Publishing. https://doi.org/10.1007/978-3-319-07443-6_39

Some lessons learnt from the study

From the study, we want to identify some guidelines for creating a sub-optimal set of citation intents that may be used for having better agreements

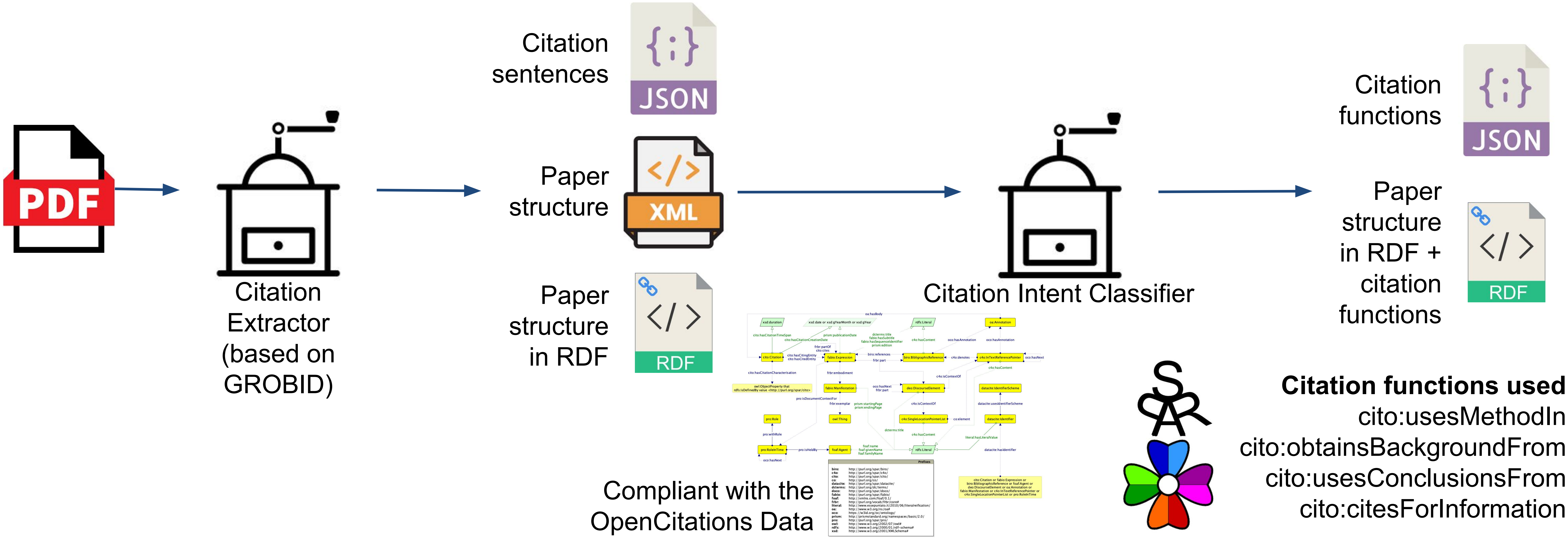
- Focus on the most used citation intents
- Provide one neutral (i.e. residual) intent

To this end, we chose a reasonable subset (from the SciCite dataset) of intents for further experimentations:

- `cito:obtainsBackgroundFrom`
- `cito:usesMethodIn`
- `cito:usesConclusionsFrom`
- `cito:citesForInformation` (neutral)

Cohan, A., Ammar, W., Van Zuylen, M., & Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. Proceedings of the 2019 Conference of the North, 3586–3596. https://doi.org/10.18653/v1/N19-1361

Need for an automatic approach



Citation Extractor (based on GROBID)

Citation sentences

Paper structure

Paper structure in RDF

Compliant with the OpenCitations Data Model (OCDM)

Citation Intent Classifier

Citation functions

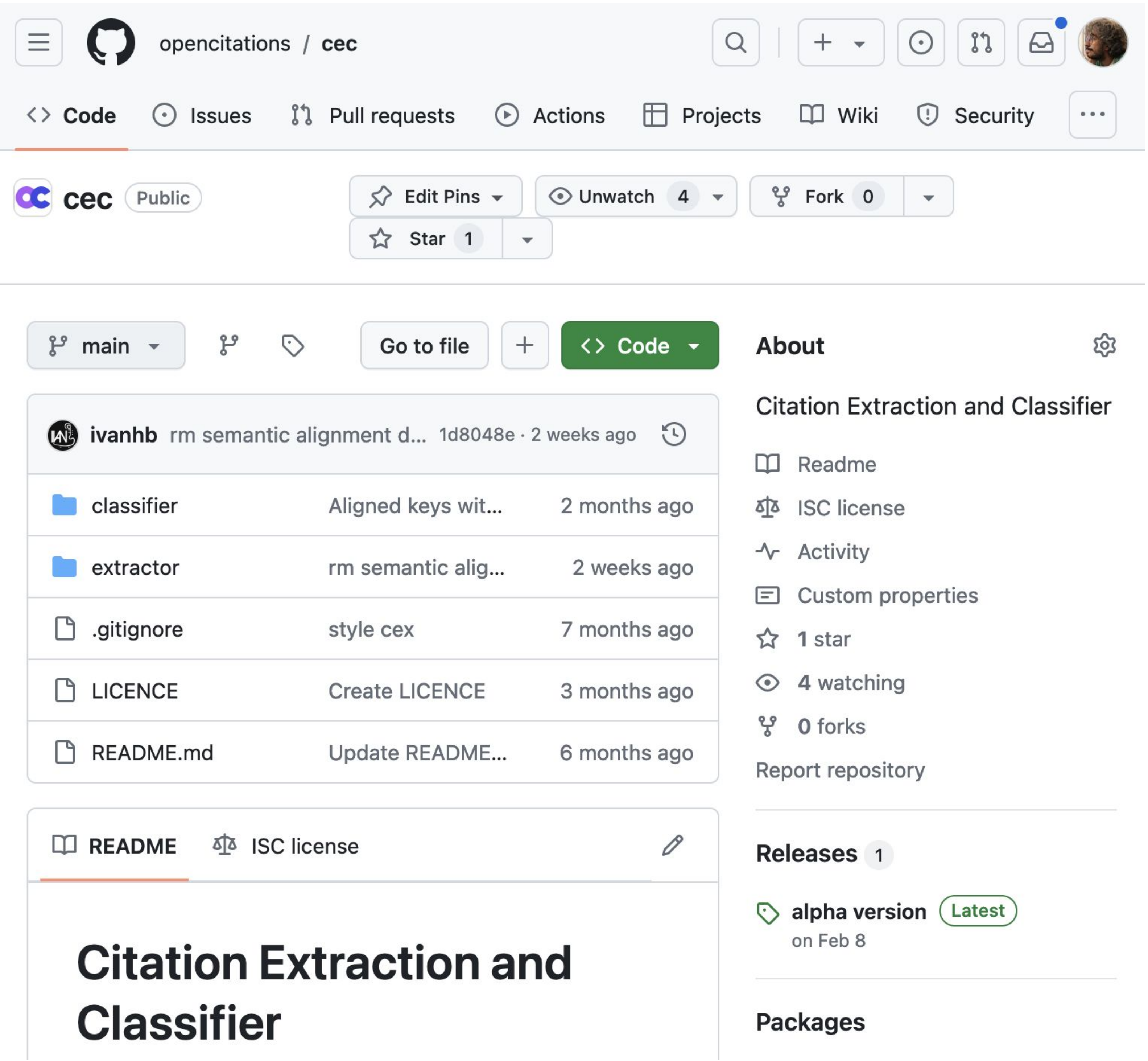
Paper structure in RDF + citation functions

Citation functions used
cito:usesMethodIn
cito:obtainsBackgroundFrom
cito:usesConclusionsFrom
cito:citesForInformation

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

The code



ivanhb rm semantic alignment d... 1d8048e · 2 weeks ago

classifier	Aligned keys wit...	2 months ago
extractor	rm semantic alig...	2 weeks ago
.gitignore	style cex	7 months ago
LICENCE	Create LICENCE	3 months ago
README.md	Update README...	6 months ago

About

Citation Extraction and Classifier

Readme
ISC license
Activity
Custom properties
1 star
4 watching
0 forks
Report repository

Releases 1

alpha version Latest
on Feb 8

Packages

Citation Extraction and Classifier

https://github.com/opencitations/cec

The Citation Extraction and Classifier is a software that performs the automatic annotation of in-text citations in academic papers provided in PDF

It is based on two steps:

1. Citation Extractor extracts basic bibliographic metadata, the bibliographic references with all its metadata marked up, the citation sentences that contain in-text reference pointers from text
2. Citation Intent Classifier classifies the semantics emerging from each citation sentence that will be used for characterising the function of the citation

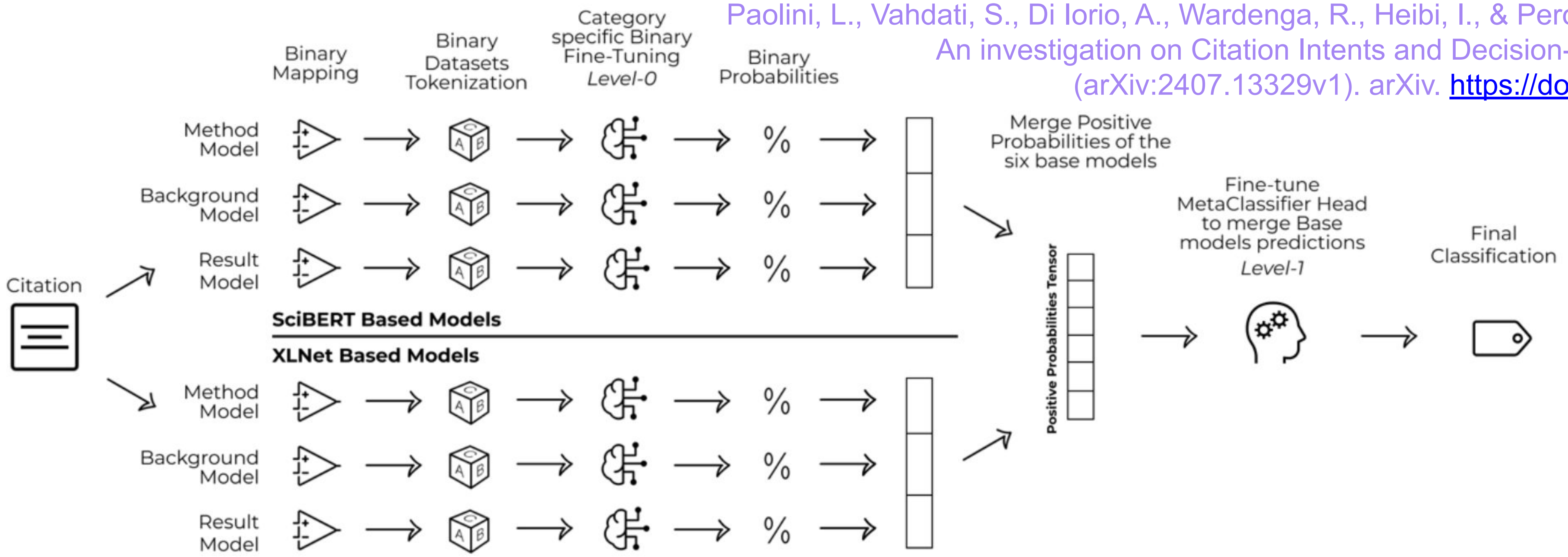
How we trained

The Citation Extractor is based on GROBID models, which have been created by training it with an additional data source we have prepared for this purpose

Pagnotta, O. (2024). CEX Project—GROBID annotation aligned Gold Standard (1.0.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10529646

The Citation Intent Classifier has been trained using the SciCite dataset, and it is based on ensemble strategies incorporating Language Models

Paolini, L., Vahdati, S., Di Iorio, A., Wardenga, R., Heibi, I., & Peroni, S. (2024). 'Why do you cite?' An investigation on Citation Intents and Decision-Making Classification Processes (arXiv:2407.13329v1). arXiv. https://doi.org/10.48550/arXiv.2407.13329


Binary Mapping
Binary Datasets Tokenization
Category specific Binary Fine-Tuning
Level-0
Binary Probabilities
Merge Positive Probabilities of the six base models
Fine-tune MetaClassifier Head to merge Base models predictions
Level-1
Final Classification
Citation
Method Model
Background Model
Result Model
SciBERT Based Models
XLNet Based Models
Method Model
Background Model
Result Model
%
Positive Probabilities Tensor

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

Live Demo

GraspOS Tools and Services Training Series
23 July 2024, 14:00-15:00 CEST

Conclusions

The Citation Extraction and Classifier is a tool developed in the context of GraspOS to extract citation information from PDF and characterise citation functions, i.e. the reason why authors cite another work, according to four different citation intents included in CiTO

We are working to extend the current code base to implement more features that will be released soon, which include:

- an in-depth documentation for installing and using the system
- REST APIs for programmatically access the Citation Extractor and the Citation Intent Classifier
- a CLI interface for running the two tools from shell
- the possibility to upload and process multiple PDFs in one run
- the export into RDF (which is not available yet)

Thank you!

Citations and their meaning

– or why we cite

OpenCitations

opencitations.net

opencitations.
hypotheses.org

@opencitations

@opencitations@
scicomm.xyz