

Towards an Annotation Standard for STEM Documents

Datasets, Benchmarks, and Spotters

Jan Frederik Schaefer Michael Kohlhase

FAU Erlangen-Nürnberg/KWARC

Conference on Intelligent Computer Mathematics (CICM)

Cambridge, UK

September 7, 2023

Natural Language Processing and Mathematical Language

- Natural language processing has benefitted from a long tradition of annotation tasks and benchmarks
- STEM documents pose problems: formulae, tables, ... *not really unicode strings*
- Why care?
 \rightsquigarrow Semantic services

Motivation: semantic services

Q 1.5 eV

↗ $1.43 \pm 0.9 \text{ eV}$

↗ $2.4 \cdot 10^{-19} \text{ J}$

Q $\sum_{k=-\infty}^{\infty} \exp(-\pi k^2)$

↗ $\sum_{n=-\infty}^{\infty} e^{-\pi n^2} = \dots$

Example from [Kri22]

equivalent to Eq. 4 can be written as follows:

$$P_{extcorr} = e^{\alpha(X_{\odot} - 1)} \times P_{meas}, \quad (5)$$

where $\alpha = 0.92103k_{\lambda}$.

Motivation: semantic services

Example from [Kri22]

equivalent to Eq. 4 can be written as follows:

$$P_{extcorr} = e^{\alpha(X_{\odot} - 1)} \times P_{meas}, \quad (5)$$

where $\alpha = 0.92103k_{\lambda}$.

"the Sun's airmass"

Amount of air in direction of sun

If relative: Divided by amount of air at zenit

Motivation: semantic services

Example from [Kri22]

equivalent to Eq. 4 can be written as follows:

$$P_{extcorr} = e^{\alpha(X_{\odot} - 1)} \times P_{meas}, \quad (5)$$

where $\alpha = 0.92103k_{\lambda}$.

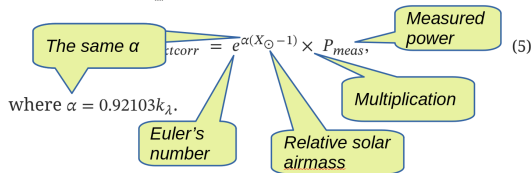
α :	<input type="text" value="0.11"/>
X_{\odot} (air mass coefficient):	<input type="text" value="1.3"/>
P_{meas} (measured power):	<input type="text" value="1 kW"/>
<input type="button" value="Compute"/>	<input type="button" value="Plot"/>

Motivation: semantic services

For all those services
we need semantic annotations!

(full formalization not necessary)

equivalent to Eq. 4 can be written as follows:



The diagram shows the equation
$$P_{atcorr} = e^{\alpha(X_{\odot}-1)} \times P_{meas}, \quad (5)$$
 with several semantic annotations in yellow callout boxes: "The same α " points to the α in the exponent; "Euler's number" points to the e base; "Relative solar airmass" points to the X_{\odot} term; "Measured power" points to the P_{meas} term; and "Multiplication" points to the \times operator. Below the equation, the text "where $\alpha = 0.92103k_{\lambda}$." is present.

where $\alpha = 0.92103k_{\lambda}$.

Motivation: semantic services

For all those services
we need semantic annotations!
(full formalization not necessary)

equivalent to Eq. 4 can be written as follows:

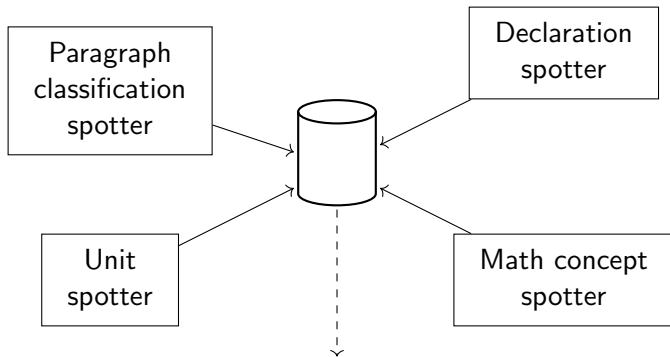
The diagram shows Equation (5) with several semantic annotations in yellow callout boxes. The equation is
$$P_{corr} = e^{\alpha(X_{\odot}-1)} \times P_{meas}, \quad (5)$$
 The annotations are:

- "The same α " pointing to the α in the exponent.
- "where $\alpha = 0.92103k_{\lambda}$." pointing to the α in the exponent.
- "Euler's number" pointing to the e in the exponent.
- "Relative solar airmass" pointing to the X_{\odot} in the exponent.
- "Measured power" pointing to P_{meas} .
- "Multiplication" pointing to the \times operator.

Authors don't provide them \rightsquigarrow We have to infer them

Accumulating semantic annotations with spotters

Spotter: specialized tool for finding a particular type of annotation



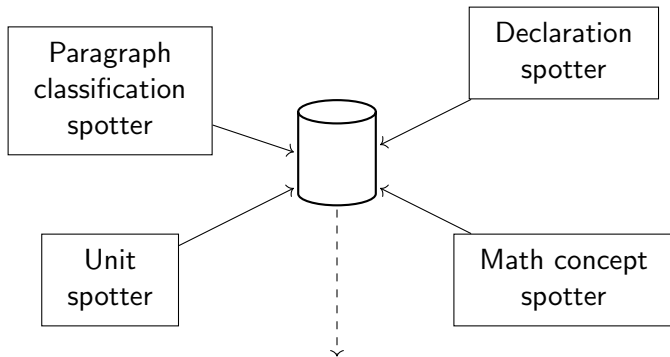
Theorem 1. *Let F be a graph with m edges and no isolated vertices. Then, for $k \geq 3$ it holds*

$$r_k(F) \leq k^{3 \cdot 2^{-1/3}} k m^{2/3} + k(2m)^{1/3} 8m.$$

Example from [JP13]

Accumulating semantic annotations with spotters

Spotter: specialized tool for finding a particular type of annotation



Theorem 1. Let F be a graph with **Fixed/frozen corpus** vertices. Then, for $k \geq 3$ it holds

$$r_k(F) \leq k^{3 \cdot 2^{-1/3}} km^{2/3} + k(2m)^{1/3} 8m.$$

Example from [JP13]

What is the problem?

① Getting a corpus:

- arXMLiv/ar5iv dataset [Gin20]
- SIGMathLing [SML]

Working with PDF is difficult

convert .tex to .html

NDA-cooperative to work around licensing issues

② Re-inventing the wheel:

- Need to obtain plaintext representation
- Need to store annotations
- Need to create manual annotations

for training/evaluation

③ Cannot re-use existing annotations/combine results:

- No agreed-upon annotation format
- Original documents modified

A new annotation standard

- Supports development of re-usable tools, datasets and benchmarks
- Uses RDF (Resource Description Framework)
 - ∃ *databases, query language (SPARQL), serialization formats*
- Based on W3C Web Annotation Recommendations

A new annotation standard

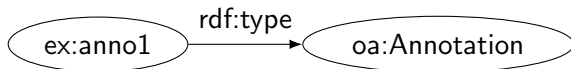
- Supports development of re-usable tools, datasets and benchmarks
 - Uses RDF (Resource Description Framework)
 - ∃ *databases, query language (SPARQL), serialization formats*
 - Based on W3C Web Annotation Recommendations
-

RDF Primer

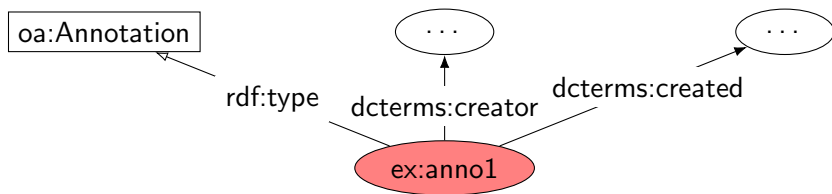
subject-predicate-object triple

`ex:anno1 rdf:type oa:Annotation`

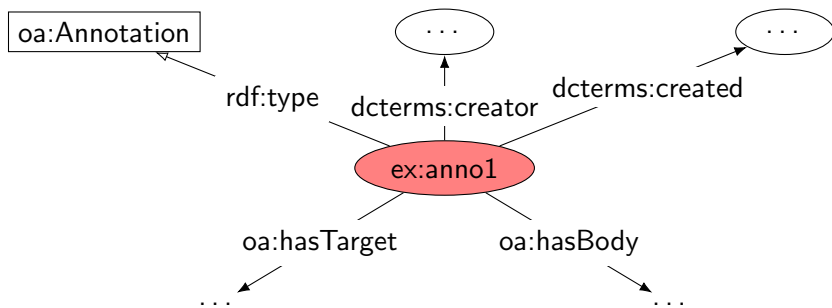
directed graph



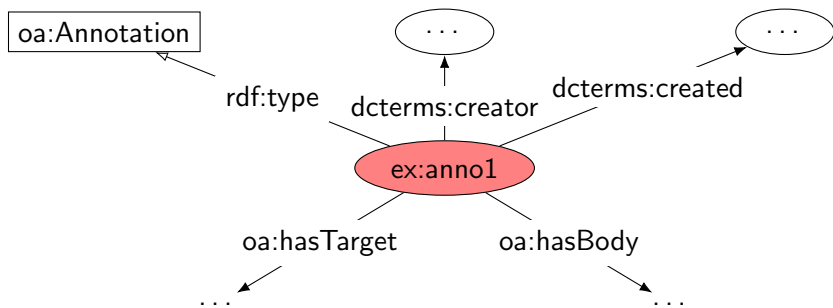
Annotation structure (following W3C Web Annotation Recommendation)



Annotation structure (following W3C Web Annotation Recommendation)



Annotation structure (following W3C Web Annotation Recommendation)



Theorem 2. Let F be a bipartite graph with m edges and isolated vertices. Then for $k > 2$ it holds

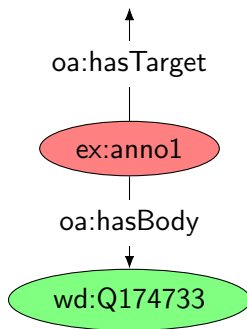
Example from [JP13]

`wd:Q174733`
(WikiData: bipartite graph)

Example annotation bodies: simple body

Example from [JP13]

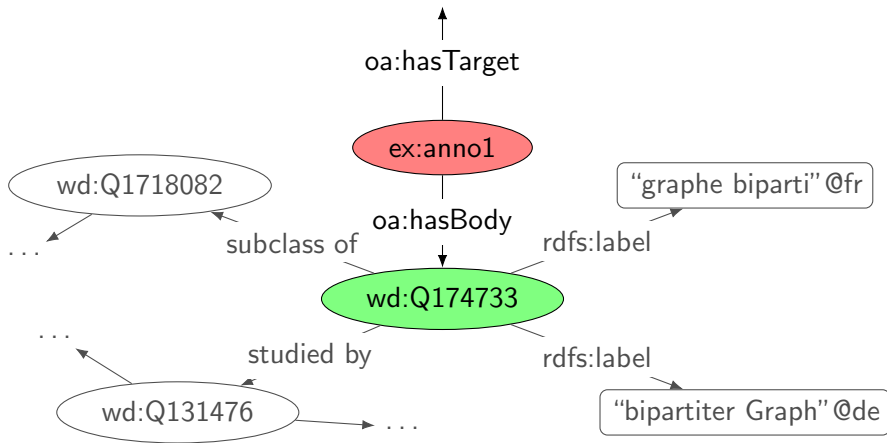
Theorem 2. *Let F be a bipartite graph with m edges and isolated vertices. Then for $k > 2$ it holds*



Example annotation bodies: simple body

Example from [JP13]

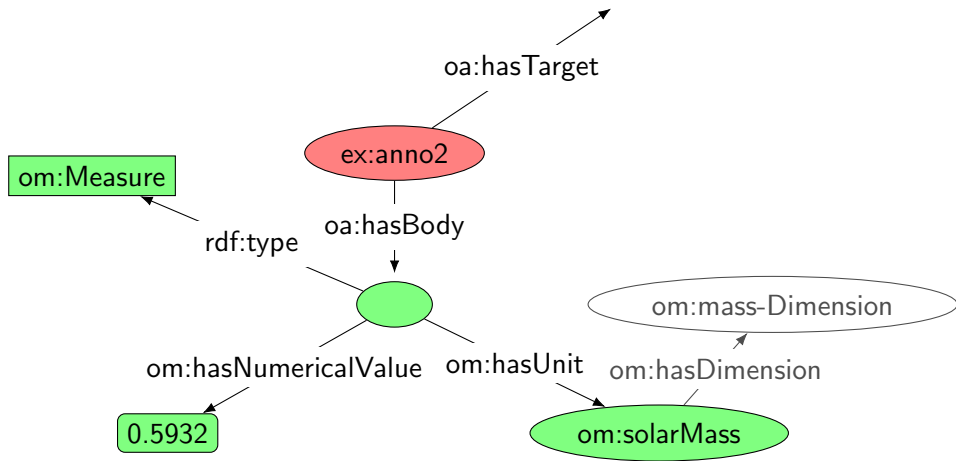
Theorem 2. Let F be a **bipartite graph** with m edges and isolated vertices. Then for $k > 2$ it holds



Example annotation bodies: complex body

Example from [AC22]

work for all the examined stellar masses. In particular, for the case of $M_{\star} = 0.5932 M_{\odot}$, the peak



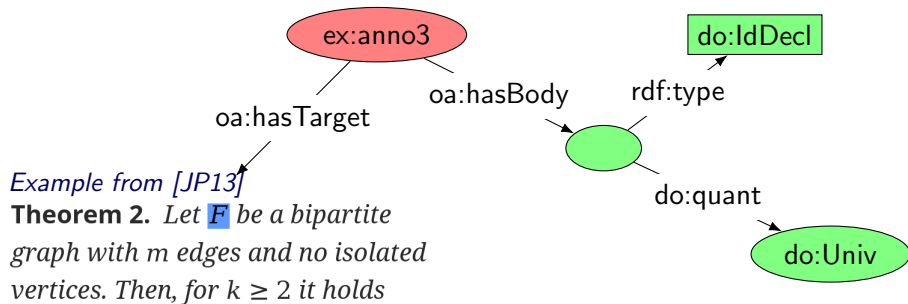
Example annotation bodies: linked bodies

Example from [JP13]

Theorem 2. Let \mathbf{F} be a bipartite graph with m edges and no isolated vertices. Then, for $k \geq 2$ it holds

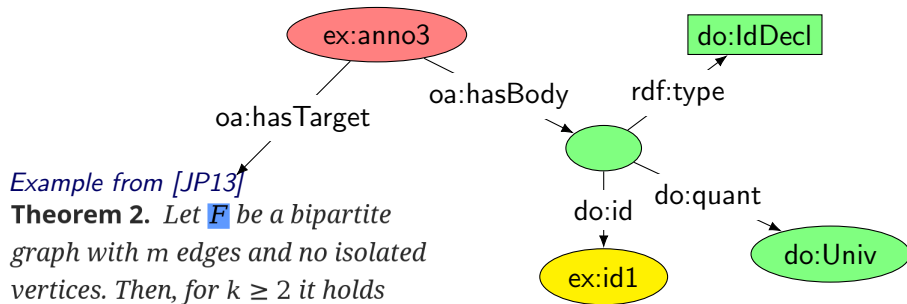
$$r_k(\mathbf{F}) \leq 2^6 m^{3/2} k^{2\sqrt{km} + 1/2}.$$

Example annotation bodies: linked bodies



$$r_k(\mathbf{F}) \leq 2^6 m^{3/2} k^{2\sqrt{km} + 1/2}.$$

Example annotation bodies: linked bodies



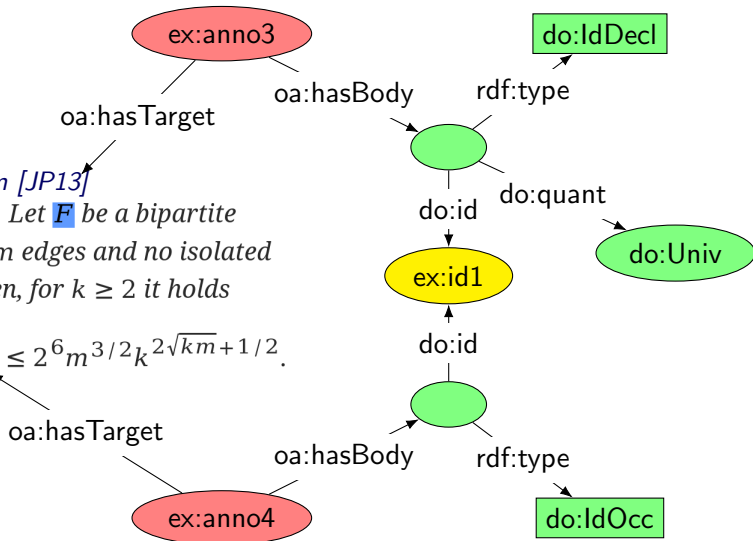
$$r_k(\mathbf{F}) \leq 2^6 m^{3/2} k^{2\sqrt{km} + 1/2}.$$

Example annotation bodies: linked bodies

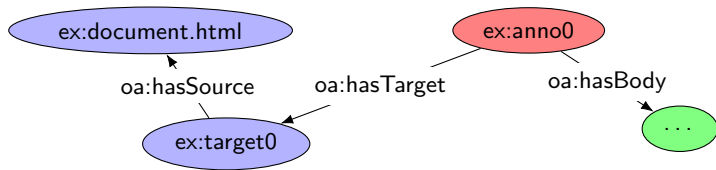
Example from [JP13]

Theorem 2. Let \mathbf{F} be a bipartite graph with m edges and no isolated vertices. Then, for $k \geq 2$ it holds

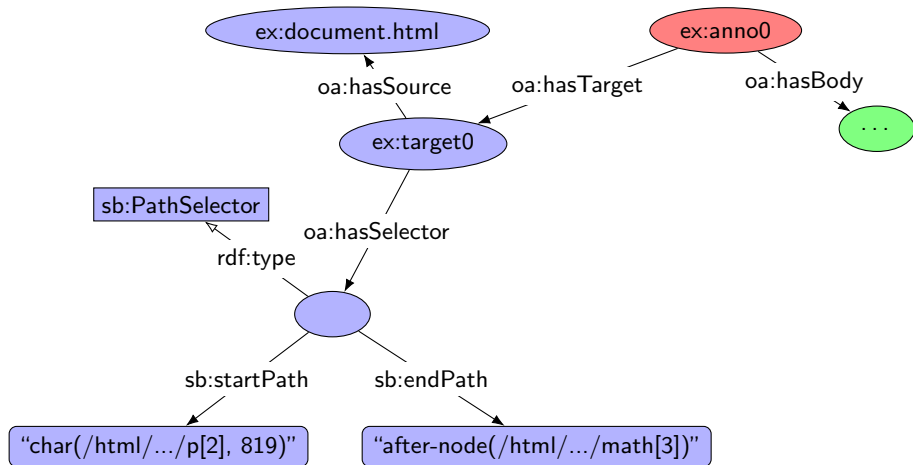
$$r_k(\mathbf{F}) \leq 2^6 m^{3/2} k^{2\sqrt{km} + 1/2}.$$



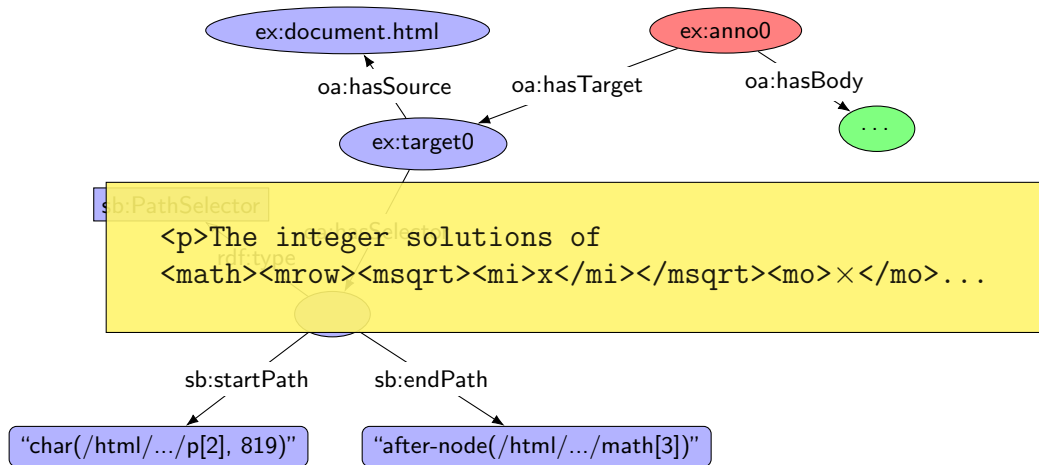
Annotation Targets



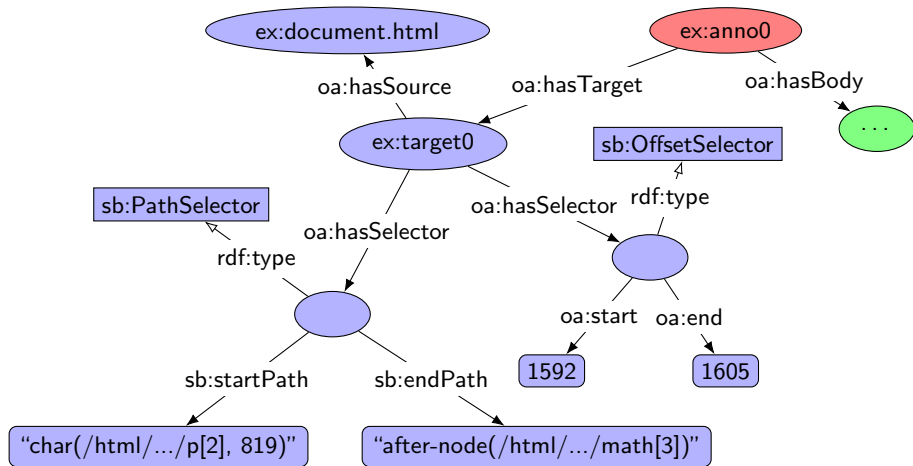
Annotation Targets



Annotation Targets



Annotation Targets



Prototype datasets and imported datasets

- Imported datasets:
 - Quantity expressions dataset [Rab17]
 - Formula grounding dataset [AMA22]
 - Paragraph classification dataset [GM20]
- Generated datasets (prototype spotters) *<https://github.com/jfschaefer/spotters>*
 - Part-of-speech tags
 - References to math concepts
 - Variable declarations
- Stats from running last two spotters:
 - 100 000 documents
 - 50 million annotations
 - 800 million triples
 - loading into triple store: several hours

Querying

"Papers about group theory that have theorems mentioning rational numbers"

```
# prefix declarations omitted for conciseness
```

```
SELECT DISTINCT ?paper WHERE {
```

```
  # make sure that ?paper is about group theory
```

```
  ?paper sb:isBasedOn/^oa:hasTarget/oa:hasBody/rdf:value arxivcat:math\..GR
```

```
  # find theorems in ?paper and look up their offsets
```

```
  ?theorem_anno oa:hasBody/rdf:value sbp:Theorem .
```

```
  ?theorem_anno oa:hasTarget [
```

```
    oa:hasSource ?paper ;
```

```
    oa:hasSelector [ a sb:OffsetSelector ; oa:start ?t_start ; oa:end ?t_end
```

```
  ] .
```

```
  # Same with mentions of rational numbers (offsets ?q_start, ?q_end)
```

```
  ?q_anno oa:hasBody/rdf:value <http://www.wikidata.org/entity/Q1244890> .
```

```
  ?q_anno oa:hasTarget [
```

```
    oa:hasSource ?paper ;
```

```
    oa:hasSelector [ a sb:OffsetSelector ; oa:start ?q_start ; oa:end ?q_end
```

```
  ] .
```

```
  # make sure that mention is inside theorem
```

```
  FILTER (?t_start < ?q_start && ?t_end > ?q_end)
```

Selling points

- Conversion to/from JSON
- Ecosystem of tools:
 - Tool for manual annotation
 - Pre-processing for NLP
 - More to come?
- Datasets
 - for evaluation
 - for comparison
 - for training
 - to build upon
- Public SPARQL endpoint

no need to learn RDF/SPARQL

MathUI workshop at 2:00 pm today

Theorem 1. Let G be a graph with m edges and no isolated vertices. Then, for $k \geq 3$ it holds

$$r_k(F) \leq k^{3 \cdot 2^{-1/3} km^{2/3} + k(2m)^{1/3}} 8m.$$

Further we study the case when F is bipartite and show an upper bound $r_k(F) \leq k^{(1+o(1))2\sqrt{mk}}$.

Theorem 2. Let G be a bipartite graph with m edges and no isolated vertices. Then, for $k \geq 2$ it holds

$$r_k(F) \leq 2^6 m^{3/2} k^{2\sqrt{km}} + 1/2.$$

Note that in the case $k = 2$, Theorem 2 is an improvement of the above mentioned result of

k

IdentifierOccurrence:

Identifier:

http://127.0.0.1:5000/docum
b88755c4-fcbe-4c7b-9f-
9e50256fb31f.anno.



An annotation standard for STEM documents

- based on semantic web technologies *RDF, SPARQL, Web Annotation Standard*
- compatible with a wide range of annotation tasks
- to create diverse, re-usable annotation datasets and benchmarks
- to develop an ecosystem of tools around that standard
- to ultimately enable the development of semantic services

active documents, formula search, . . .

References I

- [AC22] Leandro G. Althaus and Alejandro H. Corsico. “New DA white dwarf models for asteroseismology of ZZ Ceti stars”. In: *Astronomy & Astrophysics* 663 (2022), A167. DOI: 10.1051/0004-6361/202243943. URL: <https://doi.org/10.1051/0004-6361/202243943>.
- [AMA22] Takuto Asakura, Yusuke Miyao, and Akiko Aizawa. “Building Dataset for Grounding of Formulae — Annotating Coreference Relations Among Math Identifiers”. In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 4851–4858. URL: <https://aclanthology.org/2022.lrec-1.519>.

References II

- [Gin20] Deyan Ginev. *arXMLiv:2020 dataset, an HTML5 conversion of arXiv.org*. SIGMathLing – Special Interest Group on Math Linguistics. 2020. URL: <https://sigmathling.kwarc.info/resources/arxmliv-dataset-2020/>.
- [GM20] Deyan Ginev and Bruce R Miller. “Scientific Statement Classification over arXiv.org”. English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1219–1226. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.153>.
- [JP13] Kathleen Johst and Yury Person. *On the multicolor Ramsey number of a graph with m edges*. 2013. arXiv: 1311.5471 [math.CO].

References III

- [Kri22] Kevin Krisciunas. *Including Atmospheric Extinction in a Performance Evaluation of a Fixed Grid of Solar Panels*. 2022. [arXiv: 2107.02876 \[astro-ph.IM\]](#).
- [Rab17] Ullrich Rabenstein. “Meaning Extraction and Semantic Services in STEM-Documents – A case study on Quantity Expressions and Units”. Master’s Thesis. Informatik, FAU Erlangen-Nürnberg, 2017. URL: <https://gl.kwarc.info/supervision/MSc-archive/blob/master/2017/urabenstein/Rabenstein.pdf>.
- [SML] *SIGMathLing – Special Interest Group on Maths Linguistics*. URL: <http://sigmathling.kwarc.info> (visited on 12/07/2018).