

AnnoTize: A Flexible Annotation Tool for Documents with Mathematical Formulae

Lukas Panzer **Jan Frederik Schaefer**

FAU Erlangen-Nürnberg/KWARC

MathUI Workshop

CICM Conference

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

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

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

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

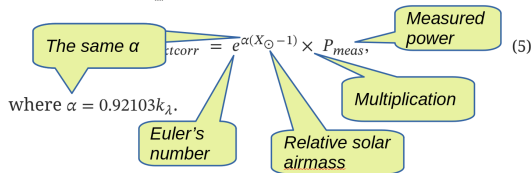
α :	<input type="text" value="0.11"/>
X_{\odot} (air mass coefficient):	<input type="text" value="1.3"/>
P_{meas} (measured power):	<input type="text" value="1 kW"/>
<input type="button" value="Compute"/>	<input type="button" value="Plot"/>

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

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

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
$$P_{corr} = e^{\alpha(X_{\odot}-1)} \times P_{meas}, \quad (5)$$
 where $\alpha = 0.92103k_{\lambda}$. Annotations include: 'The same α ' pointing to α ; 'Euler's number' pointing to e ; 'Relative solar airmass' pointing to X_{\odot} ; 'Measured power' pointing to P_{meas} ; and 'Multiplication' pointing to the \times operator.

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

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

We will need manual annotations

for evaluation and possibly training

Formulae prevent us from using the standard tools

no reasonable plaintext representation

~→ We present **AnnoTize**, a flexible annotation tool for math documents

<https://github.com/rezakul/AnnoTize>

AnnoTize

- is an annotation tool for HTML documents with a particular focus on formula support
- supports a wide range of annotation types
 - ABoSpec files for new types of declarations*
- makes the annotation process more efficient with templates

References I

- [Kri22] Kevin Krisciunas. *Including Atmospheric Extinction in a Performance Evaluation of a Fixed Grid of Solar Panels*. 2022. [arXiv: 2107.02876 \[astro-ph.IM\]](#).