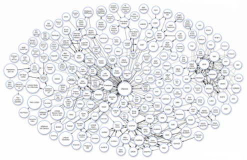
Logical Structure Analysis of Scientific Publications in Mathematics

Valery Solovyev, **Nikita Zhiltsov**

Kazan (Volga Region) Federal University, Russia

WIMS'11

Overview



- LOD Cloud has been growing at 200-300% per year since 2007*
- Prevalent domains: government (43%), geographic (22%) and life sciences (9%)
- However, it lacks data sets related to academic mathematics

*C.Bizer et al. State of the Web of Data. LDOW WWW'11

Mathematical Scholarly Papers

Essential features

- Well-structured documents
- The presence of mathematical formulae
- Peculiar vocabulary ("*mathematical vernacular*")

Research Objectives

Current study

- Specification of the document logical structure
- Methods for extracting structural elements

Long-term goals

- A large corpus of semantically annotated papers
- Semantic search of mathematical papers

Modelling the Structure of Scientific Publications

ABCDE format

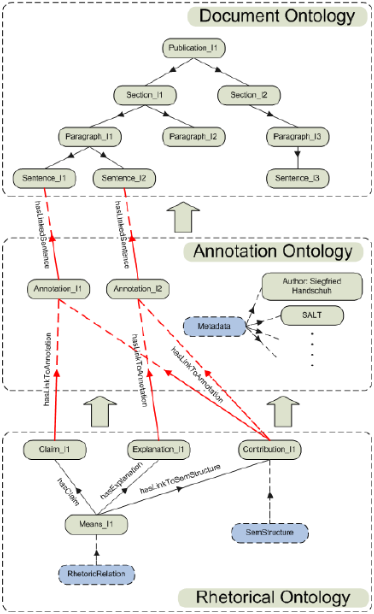
- LaTeX-based format to represent the narrative structure of proceedings and workshop contributions
- Sections:
  - **A**nnotations (Dublin Core metadata)
  - **B**ackground (e.g. description of research positioning)
  - **C**ontribution (description of the presented work)
  - **D**iscussion (e.g. comparison with other work)
  - **E**ntities (citations)

Modelling the Structure of Scientific Publications

SALT

- LaTeX-based authoring tool for generating semantically annotated PDF documents
- Three ontologies:
  - SALT Document Ontology
  - SALT Annotation Ontology
  - SALT Rhetorical Ontology

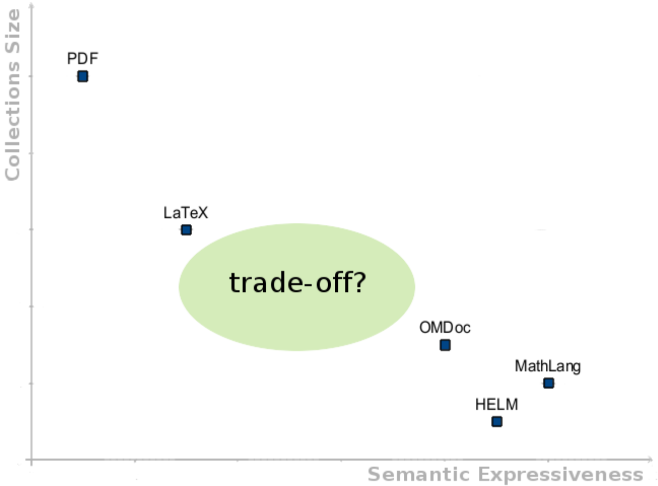
SALT Layers


Document Ontology
Publication_I1
Section_I1
Section_I2
Paragraph_I1
Paragraph_I2
Paragraph_I3
Sentence_I1
Sentence_I2
Sentence_I3
hasLinkedSentence
hasLinkedSentence
Annotation Ontology
Annotation_I1
Annotation_I2
Author: Siegfried Handschuh
SALT
Metadata
hasLinkToAnnotation
hasLinkToAnnotation
hasLinkToAnnotation
Claim_I1
Explanation_I1
Contribution_I1
hasClaim
hasExplanation
hasLinkToSemStructure
Means_I1
SemStructure
RhetoricRelation
Rhetorical Ontology

Mathematical Knowledge Representation

- Languages for formalized mathematics
  - Mizar
  - Coq
  - Isabelle
- Semiformal math languages
  - HELM ontology
  - MathLang
  - OMDoc format (+ OMDoc ontology, sTeX)
- Presentation/authoring formats
  - PDF
  - LaTeX

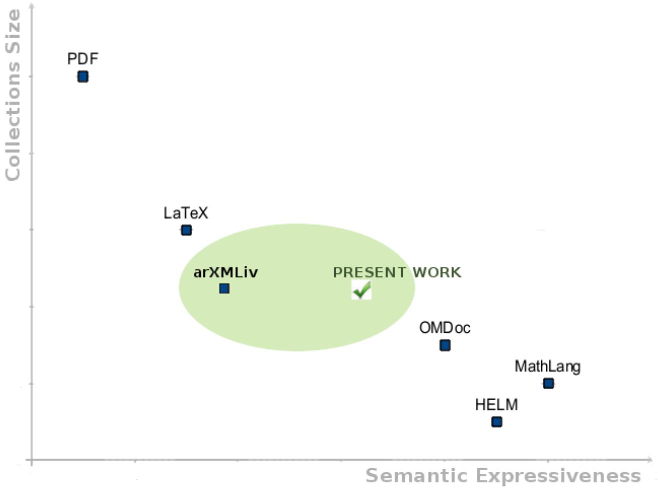
Mathematical Knowledge Representation



Trade-off Candidates

- arXMLiv format
  - XHTML+MathML
  - Marked up theorem-like elements, sections, equations
  - Automatic conversion for LaTeX documents with styles of available bindings (LaTeXML)
  - 60% of arXiv.org were converted into the format
- Present work
  - Follow the slides ⇒

Mathematical Knowledge Representation



Proposed Semantic Model

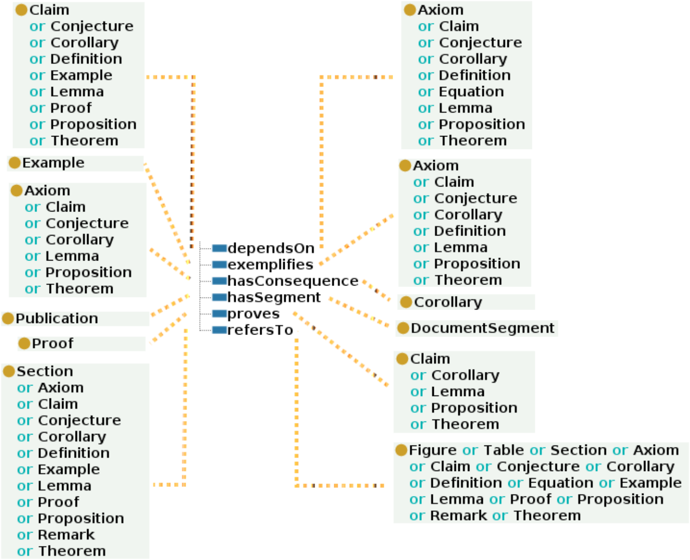
- It is an ontology that captures the structural layout of mathematical scholarly papers (as in the LaTeX markup)
- The segment represents the finest level of granularity and has the properties:
  - starting and ending positions
  - the text or math contents
  - functional role
- Select most frequent segments from sample collections of genuine papers
- Consider synonyms as one concept (e.g. *conjecture* and *hypothesis*)

Proposed Semantic Model (cont.)

- Select basic semantic relations between segments from the prior-art models
- Integration with SALT Document Ontology classes:
  - Publication
  - Section
  - Figure
  - Table

Ontology Elements

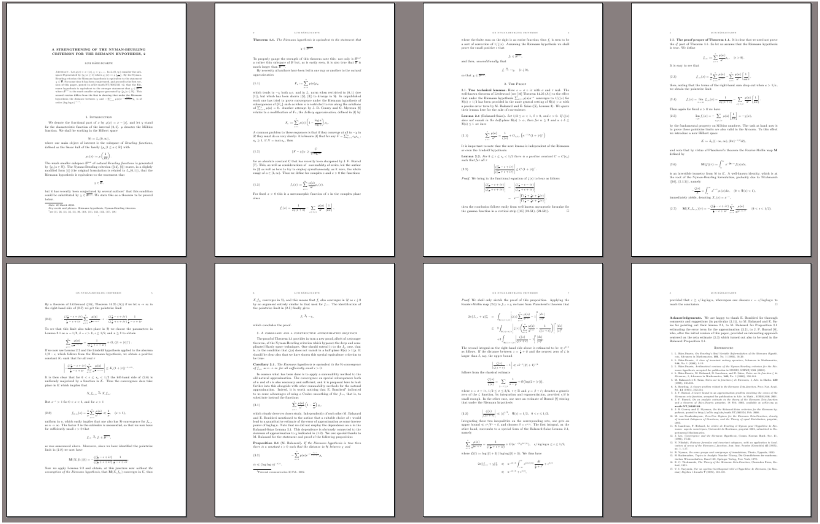
http://cll.niimm.ksu.ru/ontologies/mocassin#



Logical Structure Analysis

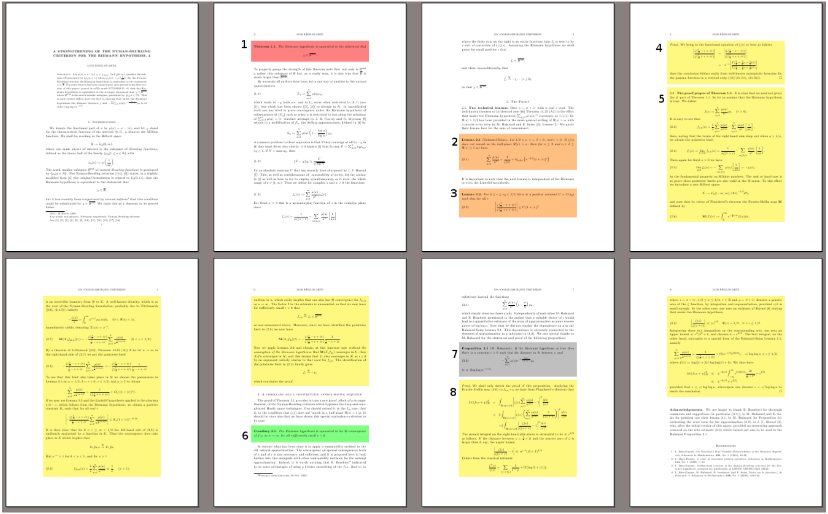
- The ontology specifies a controlled vocabulary to semantic analysis
- Two analysis tasks:
  - recognizing the types of document segments
  - recognizing the semantic relations between them

Example

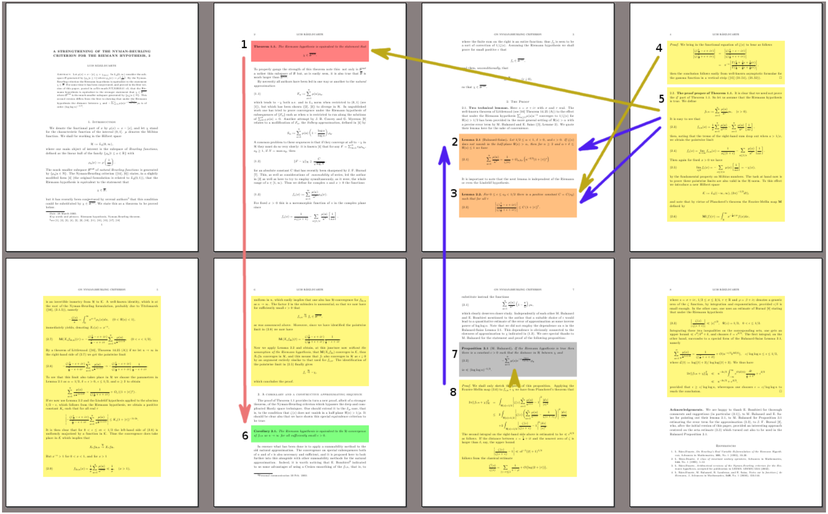

A STRENGTHENING OF THE NYMAN-BEURLING CRITERION FOR THE RIEMANN HYPOTHESIS, 2

LUIS BÁEZ-DUARTE

Example (cont.)



Example (cont.)



Recognizing the Types of Document Segments

We exploit the LaTeX markup extensively

1. Elicit a LaTeX environment
2. Associate it with a string that may be

   either the environment name

   or the environment title (if available)
3. Filter out standard formatting environments (e.g. *center*, *align*, *itemize*)
4. Compute string similarity between a string and canonical names of ontology concepts
5. Check if the found most similar concept is appropriate using a predefined threshold

Recognizing Navigational Relations

The *dependsOn* and *refersTo* relations are navigational

Assumption

Navigational relations are induced by referential sentences

Examples

- "By applying Lemma 1, we obtain ..." (dependsOn)
- "Theorem 2 provides an explicit algorithm ..." (refersTo)

Recognizing Navigational Relations

Supervised method

1. Given a segment *S*; split its text into sentences, tokenize and do POS tagging
2. Referential sentences are ones that contain the \ref command entries
3. For each sentence:
   - find mentioned segments; each of them makes a pair with *S* (type feature)
   - for each pair, compute relative positions of segments normalized by the document size (distance feature)
   - build a boolean vector for its verbs (verb feature)

Recognizing Navigational Relations (cont.)

Supervised method

Example training instance

t1	t2	d1	d2	add	...	apply	...	relation
proof	lemma	0.09	0.27	0	...	1	...	dependsOn

- Train a learning model using these features and a labeled example set
- Apply the model to classify new induced relations

Recognizing Restricted Relations

The *hasConsequence*, *exemplifies* and *proves* relations are restricted

Assumption

Restricted relations occur between consecutive segments

Recognizing Restricted Relations (cont.)

Baseline method

According to the ontology, restricted relations involve instances of three types, separately: *Corollary*, *Example* and *Proof*

1. Seek a segment of one of these types
2. Find its segments-predecessors
3. Filter out segments of inappropriate types
4. Return the closest predecessor

Experimental Setup

Collections

- 1355 papers of the “Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika” journal
- A sample of 1031 papers from arXiv.org

Implementation

An open source Java library built upon:

- LaTeX-to-XML converters
- GATE framework
- Weka
- Jena

See `http://code.google.com/p/mocassin`

Segment Recognition Evaluation

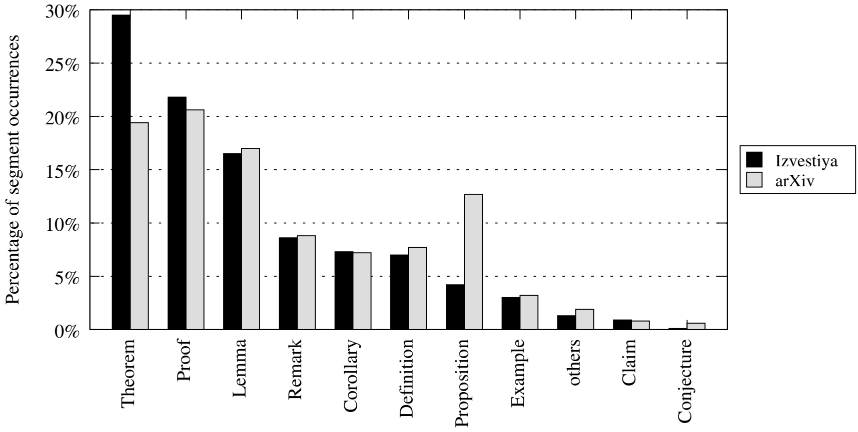
- Evaluation on the arXiv sample only
- Q-gram string matching algorithm was used
- The threshold value was optimized w.r.t. F_1-score

Type	**# of true instances**	**F_1-score**
Axiom	5	1.000
Claim	114	0.987
Conjecture	152	0.987
Corollary	1715	0.995
Definition	1838	1.000
Example	771	0.999
Lemma	4061	0.998
Proof	4943	0.997
Proposition	3052	0.999
Remark	2114	1.000
Theorem	4670	0.991
other	671	0.892

Ontology Coverage Evaluation

- Evaluation on the both entire collections ("Izvestiya" and arXiv)
- Equations are most ubiquitous segments (52% and 69%, respectively)
- The ontology covers types of 91.9% and 91.6% of segments (with SALT Section class – 99.5% and 99.6%)

Distribution of Segment Types



Evaluation of Navigational Relation Recognition

- A paper contains 51.4 (Izvestiya) and 53.9 (arXiv) referential sentences on the average
- 243 referential sentences were randomly selected and manually annotated
- 95% were true navigational relations
- A decision tree learner (C4.5) was trained
- The results were from 10-fold cross validation

Features	**Accuracy**	**F_1-score** refersTo	**F_1-score** dependsOn
type	0.663	0.566	0.752
type+distance	0.658	0.663	0.704
type+verb	0.704	0.653	0.770
type + distance + verb	**0.741**	**0.744**	**0.772**

A Cloud of Frequent Verbs

Evaluation of Restricted Relation Recognition

- Evaluation on the arXiv sample only
- 10% of the documents which contain certain segments were randomly selected
- For each such a segment, corresponding relations were annotated manually
- Known issues: imported corollaries and examples for arbitrary text fragments

Relation	**# of instances**	F_1**-score**
hasConsequence	178	0.687
exemplifies	62	0.613
proves	216	0.954

Conclusion on Evaluation

- The ontology covers the largest part of the logical structure and appears to be feasible for automatic extraction methods
- The task of segment type recognition has been accomplished
- The method for recognizing navigational relations establishes ground truth, however, a large-scale evaluation and learning model selection are required
- The baseline method for recognizing restricted relations must be improved by leveraging additional information (discussed in the paper!)

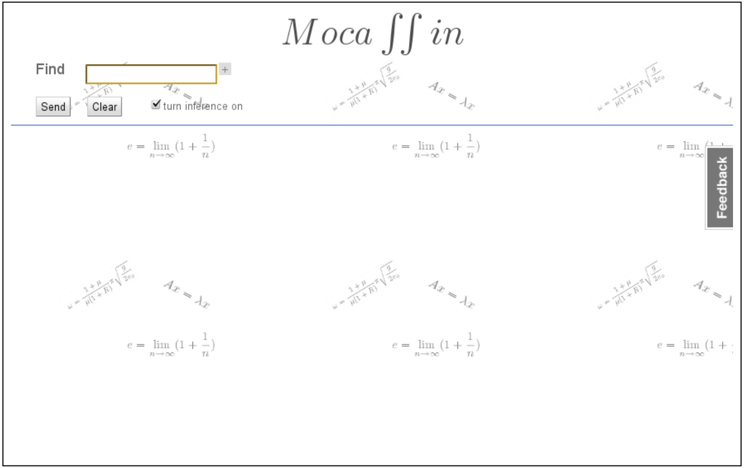
Prototype

A prototype:

- demonstrates our ongoing research on semantic search of mathematical papers
- incorporates the logical structure analysis methods
- is integrated with arXiv API
- enables enhanced search for arXiv papers and visualization of their logical structure
- publishes the semantic index as Linked Data via SPARQL endpoint

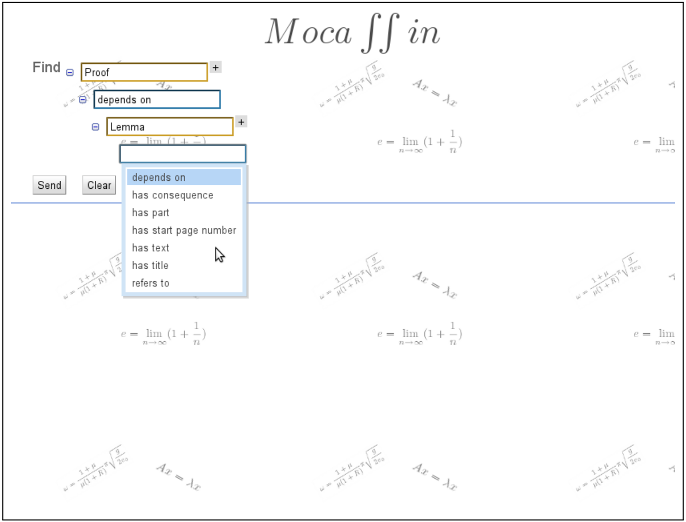
Search Interface

http://cll.niimm.ksu.ru/mocassin



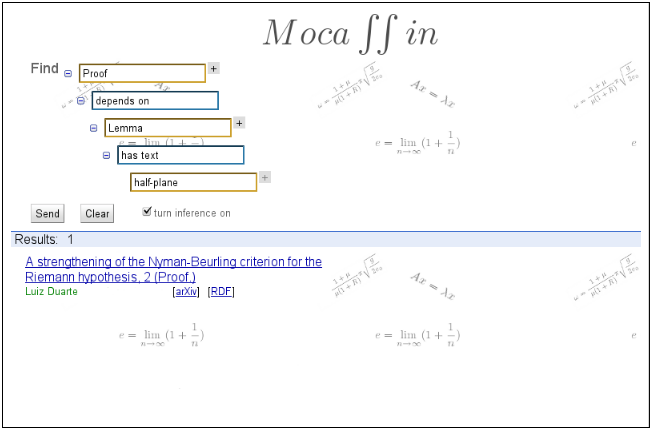
Formulating a Query

http://cll.niimm.ksu.ru/mocassin



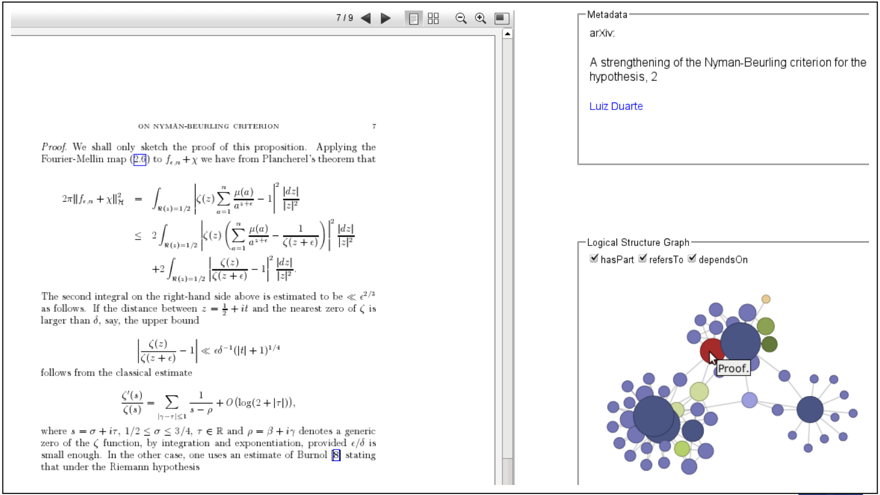
Search Results

http://cll.niimm.ksu.ru/mocassin



Preview a Search Result

http://cll.niimm.ksu.ru/mocassin



ON NYMAN-BEURLING CRITERION 7

Proof. We shall only sketch the proof of this proposition. Applying the Fourier-Mellin map (2.6) to $f_{\epsilon,n} + \chi$ we have from Plancherel's theorem that

$$\begin{aligned} 2\pi \|f_{\epsilon,n} + \chi\|^2_{\mathcal{H}} &= \int_{\Re(z)=1/2} \left|\zeta(z)\sum_{a=1}^{n}\frac{\mu(a)}{a^{z+\epsilon}} - 1\right|^2 \frac{|dz|}{|z|^2} \\ &\leq 2\int_{\Re(z)=1/2} \left|\zeta(z)\left(\sum_{a=1}^{n}\frac{\mu(a)}{a^{z+\epsilon}} - \frac{1}{\zeta(z+\epsilon)}\right)\right|^2 \frac{|dz|}{|z|^2} \\ &\quad +2\int_{\Re(z)=1/2} \left|\frac{\zeta(z)}{\zeta(z+\epsilon)} - 1\right|^2 \frac{|dz|}{|z|^2}. \end{aligned}$$

The second integral on the right-hand side above is estimated to be $\ll \epsilon^{2/3}$ as follows. If the distance between $z = \frac{1}{2} + it$ and the nearest zero of ζ is larger than δ, say, the upper bound

$$\left|\frac{\zeta(z)}{\zeta(z+\epsilon)} - 1\right| \ll \epsilon\delta^{-1}(|t|+1)^{1/4}$$

follows from the classical estimate

$$\frac{\zeta'(s)}{\zeta(s)} = \sum_{|\gamma-\tau|\leq 1} \frac{1}{s-\rho} + O(\log(2+|\tau|)),$$

where $s = \sigma + i\tau$, $1/2 \leq \sigma \leq 3/4$, $\tau \in \mathbb{R}$ and $\rho = \beta + i\gamma$ denotes a generic zero of the ζ function, by integration and exponentiation, provided ϵ/δ is small enough. In the other case, one uses an estimate of Burnol [8] stating that under the Riemann hypothesis

Summary

- The proposed approach aims to analyze the structure of mathematical scholarly papers in an automatic way
- Our ontology provides a controlled vocabulary for analysis
- The methods elicit document segments in terms of the ontology
- The extracted semantic graph can be used for:
  - discovering important document parts
  - semantic search of theoretical results

Thanks for your attention!
Questions?