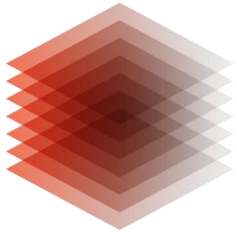
LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY



TIB

Knowledge organization systems in mathematics and in libraries

Dr. Anna Kasprzik,
Salzburg, 11. September 2017
ÖMG-Kongress / DMV-Jahrestagung

Agenda

1. A brief history of subject cataloguing, electronical data processing, and the Semantic Web

2. Knowledge organization systems in mathematics, in libraries, and beyond
   - multilinguality and multiple data sources
   - thesauri and ontologies

3. Semantic annotation – a dying trade? Suggestions for a fruitful symbiosis?

FID Mathematik –
„Specialized Information Service for Mathematics”

SUB Göttingen, TIB Hannover, L3S Hannover, FAU Erlangen

Beyond the endeavour to provide state-of-the-art research in mathematics with the **necessary information resources**, various tasks, such as:

- Digitization and storage of **mathematical legacy documents**
  → *SUB, Katharina Habermann*

- Citability and storage of **new media formats** for mathematics: software, audio-visual material, blog entries on web platforms
  → *L3S & TIB, Helge Holzmann*

- Mathematical **vocabulary**, translation and wikification services
  → *TIB & FAU, will make an appearance in this talk*

A brief history of subject cataloguing

In the beginning there was classificatory cataloguing...

- First systematic catalogues came about in the middle ages (for inventory purposes)

- These were developed further during the 18th century → **„Realkatalog“** SUB Göttingen (started in 1738)

www.phillipmartin.info

Beginning of 20th century

Number of documents starts to grow exponentially so that systematic subject cataloguing becomes indispensable

→ „**Referateorgane**“, such as:

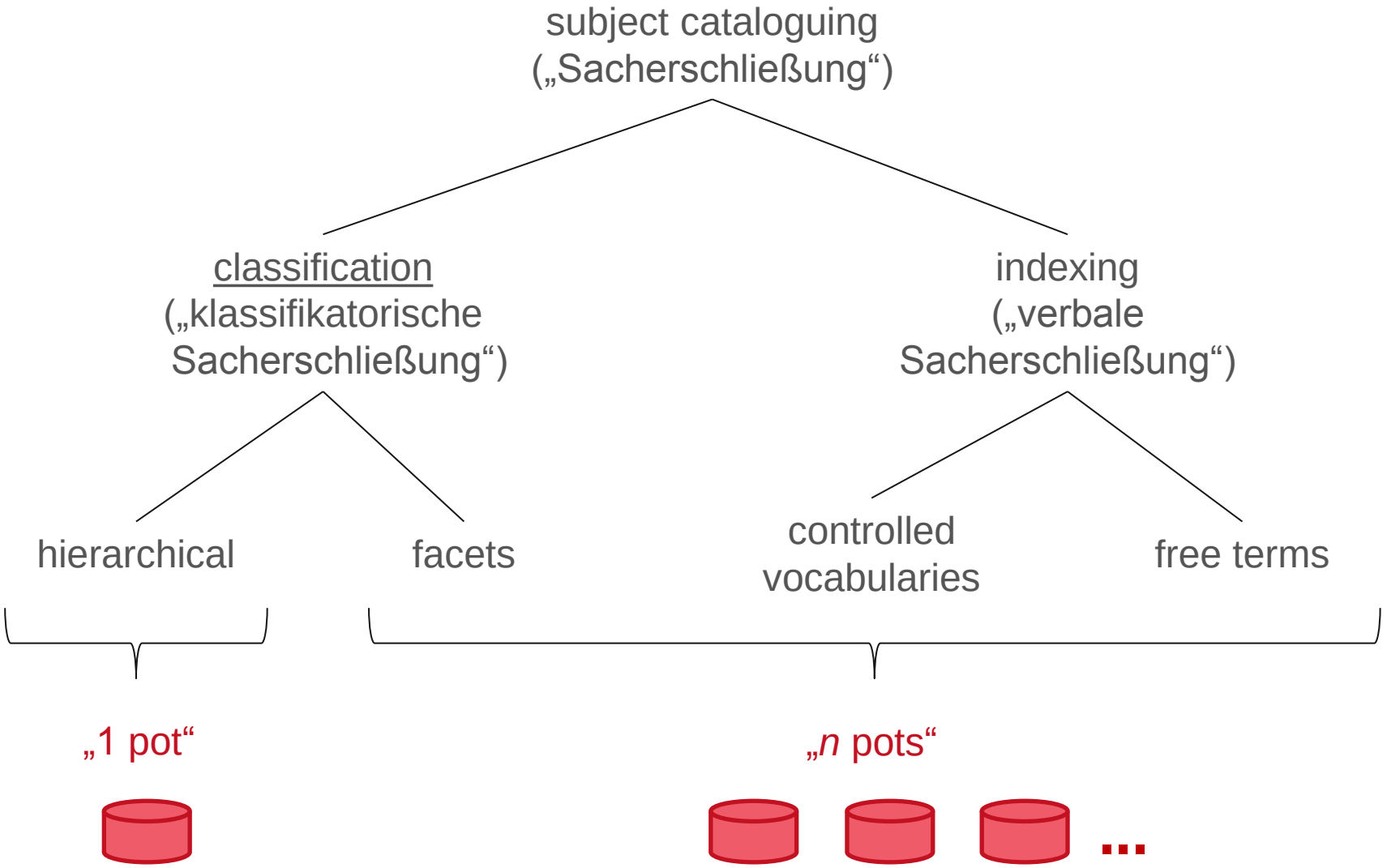
- Jahrbuch über die Fortschritte der Mathematik (1868–1942)
- Zentralblatt MATH (since 1931)
- Mathematical Reviews (since 1940)
- Referatiwnij Schurnal Matematika (since 1945), Russian

Further increase in publications numbers: the emergence of indexing

Coordinated standardization efforts in Germany:

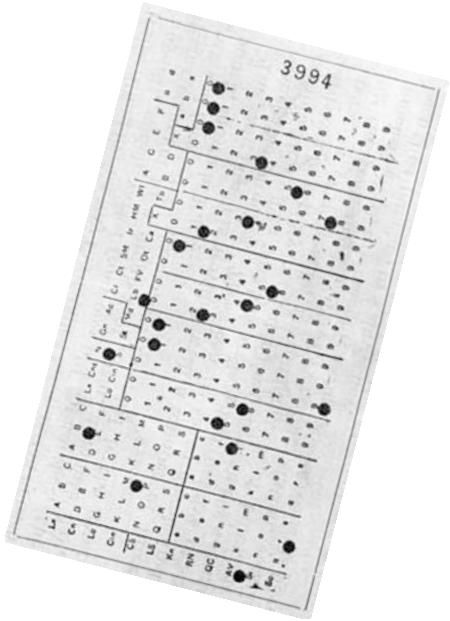
- 1980s: „Regeln für den Schlagwortkatalog“
- 1990: „Schlagwortnormdatei“
- 2012: „Personennamendatei“ (PND), „Gemeinsame Körperschaftsdatei“ (GKD), „Schlagwortnormdatei“ (SWD) and the „Einheitssachtitel-Datei des Deutschen Musikarchivs“ (DMA-EST-Datei) are merged into the **GND** („Gemeinsame Normdatei“; German Authority File for subject headings)

Subject cataloguing

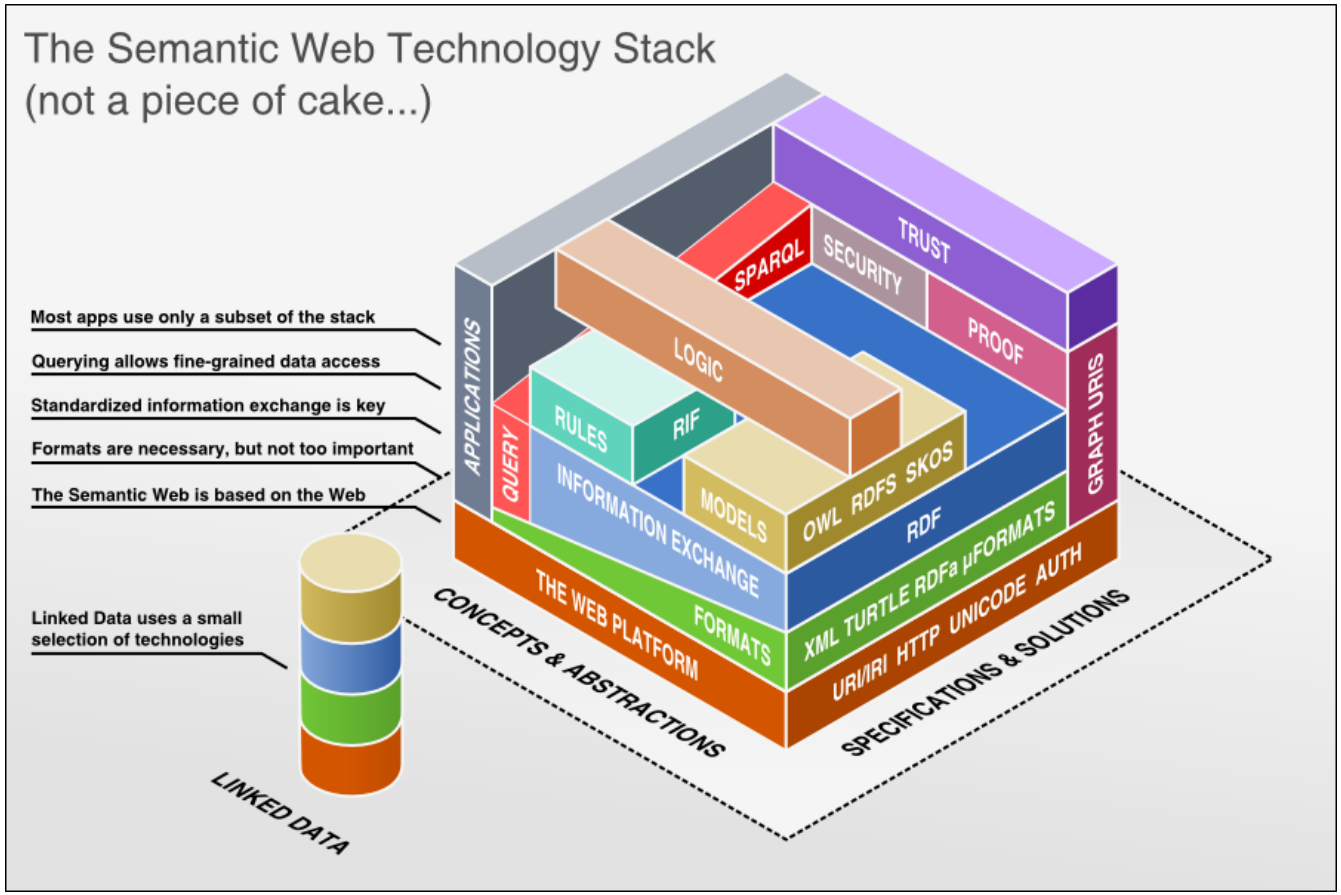


Automation of (meta)data processing in libraries

- 1936: University of Texas begins using a punch card system to manage library circulation
- 1960s: **MARC (MAchine-Readable Cataloging)** standards – library automation is born
- 1980s: emergence of **integrated library systems (ILSs)** with modules for
  - acquisition
  - circulation
  - cataloguing
- 1990s: **OPACs (Online Public Access Catalogues)** evolve
- 2010s: rise of **cloud-based solutions**



The Semantic Web



... for mathematics ?

... in libraries ?

Thesauri and ontologies (Semantic Web)

Thesauri

- natural language based
- hierarchy/network of concepts with labels and with additional semantic relations between them
- standard format: e.g., SKOS and extensions (SKOS-XL, iso-thes)

Ontologies

- logic based
- classes, relations, properties, and rules
- standard format: e.g., OWL

Hybrids/blended versions are possible (various degrees of formality)

Semantic relations – I

Two terms are synonyms if they refer to the same concept.

„Pferd“ – „Gaul“

A term is a homonym if it can refer to more than one concept.

„Bank <Sitzgelegenheit>“ – „Bank <Finanzinstitut>“

Term t_1 is hyperonym to term t_2 (t_2 is hyponym to t_1) if term t_1 refers to a concept that is a superset of a concept term t_2 refers to.

„Säugetier“ – „Pferd“

„Finanzinstitut“ – ?? – „Bank <Sitzgelegenheit>“

Semantic relations – I

Two terms are synonyms if they refer to the same concept.

„Pferd“ – „Gaul“

A term is a homonym if it can refer to more than one concept.

„Bank <Sitzgelegenheit>“ – „Bank <Finanzinstitut>“

Term t_1 is hyperonym to term t_2 (t_2 is hyponym to t_1) if term t_1 refers to a concept that is a superset of a concept term t_2 refers to.

„Säugetier“ – „Pferd“

„weißer Gaul“ — syn — „weißes Pferd“ — syn — „Schimmel“ — syn — „Schimmelpilzbefall“

Semantic relations – II

Two terms are antonyms if they refer to concepts that exclude each other and that are perceived as two extremes of a spectrum.

„Hitze“ – „Kälte“

Term t_2 is meronym to t_1 (t_1 is holonym to term t_2) if term t_2 refers to a concept of which every individual is part of some individual from a concept term t_1 refers to.

„Nase“ – „Gesicht“

„Marktplatz“ – ?? – „Bundesland“

„Auto“ – „Rad“ – „Schraube“

??

GND and VIAF

GND – a high-quality semantic network? Some deficiencies:

- inaccurate relations, very few relations, missing hyperonyms
  → hierarchy is incomplete, disconnected
- vocabulary not up-to-date, terms from state-of-the-art research are missing, missing information sources, little multilinguality
- rudimentary LOD versions of GND (SKOS, JSON-LD) exist but the specificity of superordinate relations („obal“, „obin“, „obpa“) gets lost

VIAF (Virtual International Authority File)

- corresponding records from different national authority files are linked by a clustering algorithm that is run every month
  → clusters and the range of VIAF IDs are not stable!

FID AP 5.1 – we intend to:

- establish high-quality concordances between the GND and classifications such as the MSC (see next slide) and DDC (Dewey Decimal Classification), for free reuse
- subdivide the GND classification system for mathematics and statistics into a hierarchy (→ a MSC-oriented backbone)
- update the mathematical vocabulary by enriching it with terms from state-of-the-art mathematical research
- interlink the GND with vocabularies from other languages
- interlink the GND with other KOS

MSC

- The **Mathematics Subject Classification (MSC)** is an alphanumerical classification scheme collaboratively produced by staff of, and based on the coverage of, the two major mathematical reviewing databases, Mathematical Reviews and Zentralblatt MATH.
- hierarchical scheme, three levels (“53”: differential geometry; “53A”: classical differential geometry; “53A45”: vector and tensor analysis)
- used in numerous contexts (journals, arXiv, recommender systems)
- was created in the 1960s, has been revised several times
- current version: MSC2010
- currently: collection of suggestions for the MSC2020 revision (https://msc2020.org/); refinement of levels 2 and 3

https://en.wikipedia.org/wiki/Mathematics_Subject_Classification

Multilinguality: Glossaries, WordNet

WordNet

- large lexical database of English, including word forms
- synsets (groupings of terms) express distinct concepts
- relations: conceptual-semantic, lexical, cross-POS
- disambiguation of words in close proximity

BabelNet

- multilingual lexicalized semantic network and ontology synsets
- was automatically created by linking Wikipedia to WordNet

[C]

☐ 全選	出處/學術領域	英文詞彙	中文詞彙	INFO
☐ 1	學術名詞 數學名詞	binomial	二項式	ℹ
☐ 2	學術名詞 數學名詞	binomial coefficient	二項式係數	ℹ
☐ 3	學術名詞 數學名詞	binomial correlation	二項相關	ℹ

Wikipedia, DBpedia, Wikidata ...

Wikipedia

- online encyclopedia, information is displayed in texts and information boxes, enriched with links to external content
- first language: English

DBpedia

- extraction of information from Wikipedia infoboxes into machine-readable RDF triples that allow for querying

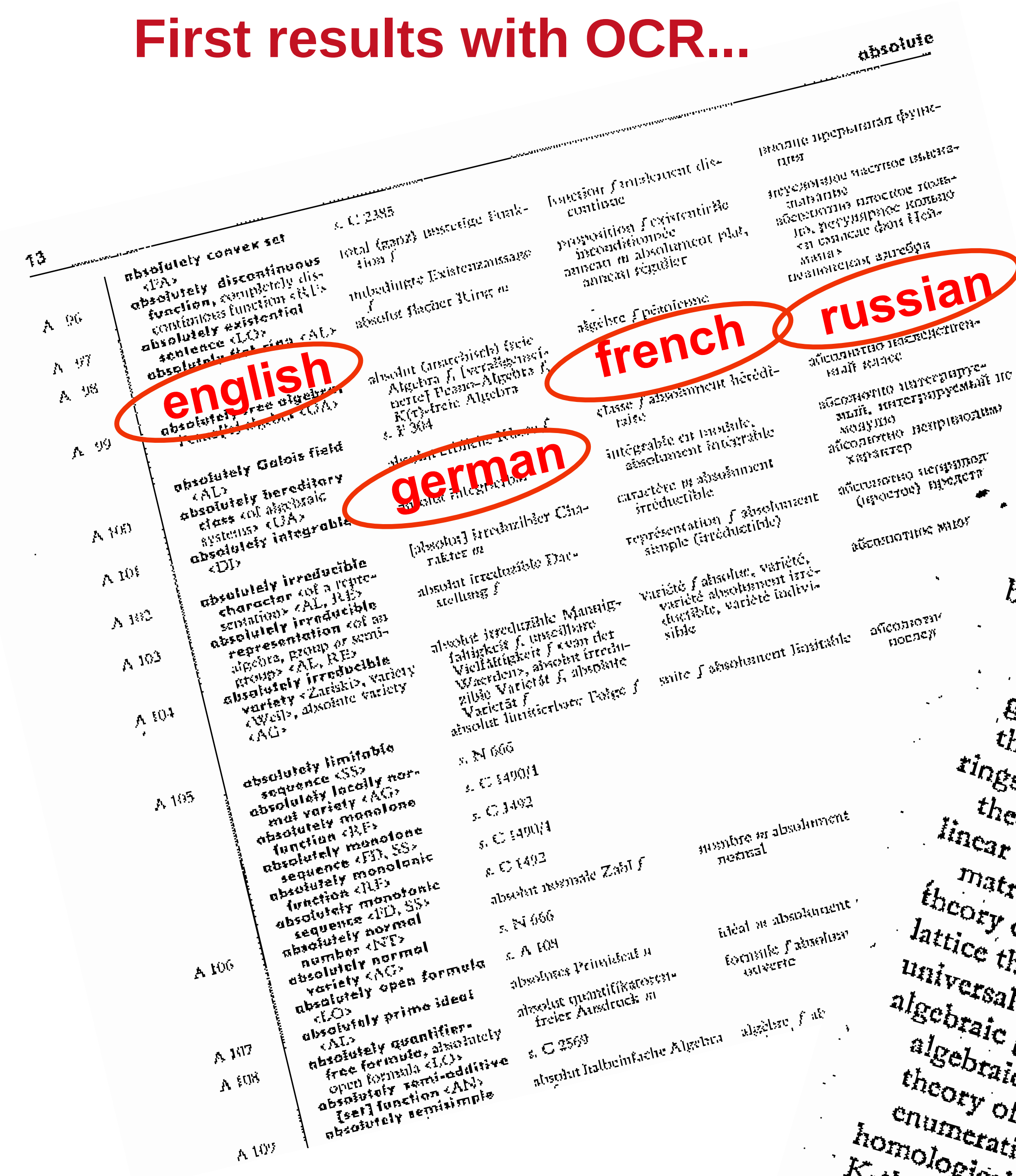
Wikidata

- (manually curated) machine-readable facts together with their sources that allow for example for the creation of infoboxes
- multinational and multilingual from the start

FID AP 5.2 – we intend to:

- digitize out-of-print dictionaries and extract semantic content (German, English, French, Russian)
- provide an online glossary for mathematics with content from dictionaries, Wikipedia and other online encyclopediae
- build a wikifier service for text enrichment
- develop a prototypical translation service for scientific texts containing mathematical research results (for LaTeX/MathML sources, e.g. from arXiv.org or mathnet.ru)
- provide a search service for formulae in online texts (see http://search.mathweb.org/)

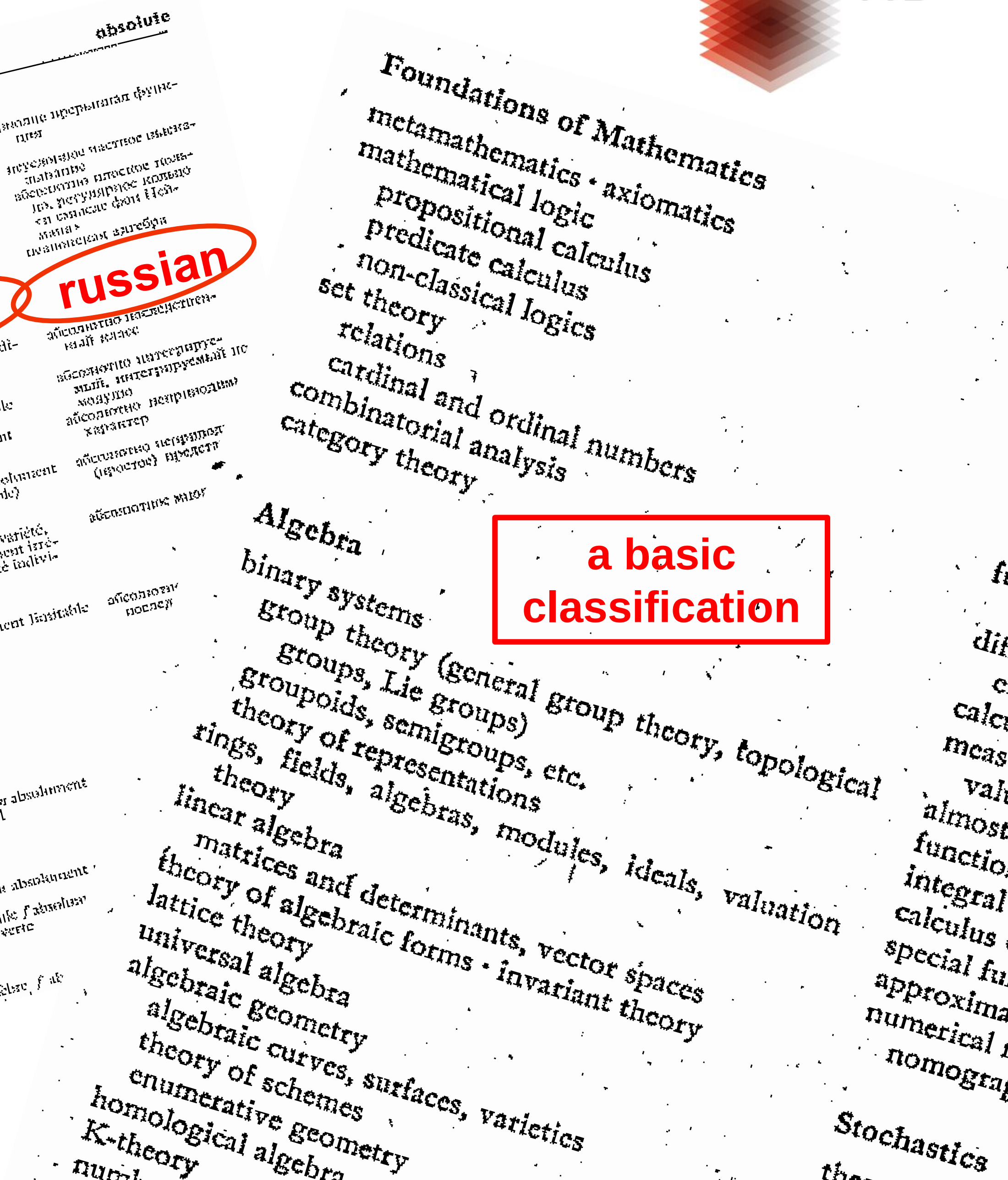
First results with OCR...



absolute

13

A 96 absolutely convex set
absolutely discontinuous function
A 97 absolutely existential sentence
A 98 absolutely free algebra
A 99
A 100 absolutely Galois field
absolutely hereditary class
A 101 absolutely integrable
A 102 absolutely irreducible character
A 103 absolutely irreducible representation
A 104 absolutely irreducible variety
A 105 absolutely limitable sequence
absolutely locally normal variety
absolutely monotone function
absolutely monotone sequence
absolutely monotonic function
absolutely monotonic sequence
absolutely normal number
A 106 absolutely normal variety
absolutely open formula
absolutely prime ideal
A 107 absolutely quantifier-free formula
A 108 absolutely semi-additive [set] function
A 109 absolutely semisimple



Foundations of Mathematics
metamathematics · axiomatics
mathematical logic
propositional calculus
predicate calculus
non-classical logics
set theory
relations
cardinal and ordinal numbers
combinatorial analysis
category theory

Algebra
binary systems
group theory (general group theory, topological groups, Lie groups)
groupoids, semigroups, etc.
theory of representations
rings, fields, algebras, modules, ideals, valuation theory
linear algebra
matrices and determinants, vector spaces
theory of algebraic forms · invariant theory
lattice theory
universal algebra
algebraic geometry
algebraic curves, surfaces, varieties
theory of schemes
enumerative geometry
homological algebra
K-theory

calculus
almost
functio
integral
calculus
special fu
approxima
numerical
nomograp

Stochastics

An alarming (?) prophecy (blog post) I

„Unsupervised machine learning has made significant strides in the past year or two, and it has become possible to extract facts from unstructured data“

„With big data’s distributed computing horsepower, semantic metadata has become just another form of data that needs to be discovered or machine generated. “

„I’m not saying semantic web techniques aren’t being used and aren’t useful, but they’ve never really become mainstream […] – they’re a sidetrack, one that tends [to favor] the permanent data world of archivists and librarians. That’s a very important place, but it’s not where the data volumes are or where the money’s being made.“

https://www.quora.com/Is-interest-in-the-semantic-web-declining-and-if-so-why-e-g-See-SPARQL-OWL-Semantic-Inference-Semantic-Reasoning-etc

An alarming (?) prophecy (blog post) I

„Unsupervised machine learning has made significant strides in the past year or two, and it has become possible to extract facts from unstructured data“

„With big data’s distributed computing horsepower, semantic metadata has become just another form of data that needs to be discovered or machine generated. “

„I’m not saying semantic ’t being used and aren’t useful, but they’ve never tream […] – they’re a sidetrack, one that tends ent data world of archivists and librarians. nt place, but it’s not where the data volumes a ey’s being made.“

https://www.quora.com/Is-interest-in-the-s hy-e-g-See-SPARQL-OWL-Semantic-Inference-Semantic-Reasoning-etc

An alarming (?) prophecy (blog post) II

„A machine-assisted process can help with entity and relationship extraction and then also ontology generation. […] Today's semantic web techniques need to work behind what machines already do and complement it. […] Hand-built ontologies only have limited relevance in a big data world."

<u>Questions (for libraries and for the semantization of mathematic objects / structures, respectively):</u>

- Do we still need hand-built semantic structures?
- Which parts of semantization can be automated?
- What would be the relevant aspects of a fruitful interaction between intellectual and automated methods?

DISCUSSION…

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

TIB

Thank you for your input.

Contact
Dr. Anna Kasprzik
T +49 511 762-14219, anna.kasprzik@tib.eu