Appendix H: Developments relevant for building a semantic collection of mathematical theorems

The Sloan Foundation funded eCF project was an exploratory project devoted to collecting and semantically representing mathematical theorems from continued fraction theory as described in detail in Appendix I. While the successful eCF project was limited in scope, it highlighted both some of the challenges that will be faced as well as some of the myriad benefits that the semantic tagging of all of mathematics will bring to human exploration, interaction, and extension of knowledge in the mathematical sciences, and in fact to all of science as a whole.

In this Appendix we outline significant other developments that can be considered precursors of building a truly semantic mathematical heritage library. To put the development of an extensive collection of semantically encoded mathematical theorems into proper context, we first review a number of relevant historical efforts. In particular, the main historic threads that are of relevance for the development, deployment, maintenance, and application of a future semantic representation of mathematics are:

1) "Classic" projects that accumulate and unify mathematical knowledge (for mathematics in general and for specialized topics; historically they were mainly realized as book series);

2) With the advance of the internet, the development of websites of accumulated mathematical knowledge;

3) With the maturing of the internet, the development of websites with semantically marked mathematical material;

4) The evolution of mathematical research projects that are carried out by many mathematicians, transforming mathematics from an endeavor carried out mainly individually to a collaborative one [183], [184], [182];

5) The development of symbolic computation software [185] and general mathematical problem-solving environments that have built-in collections of mathematical data;

6) The rise of computer-based mathematical theorem provers [186], [206],

7) The classification of mathematical research and the development of unifying and standardizing mathematical vocabularies; and

8) Investigations into the language of mathematics itself.

Each of these development threads has been worked on by many mathematicians, and the following selection highlights some developments especially relevant for a semantic mathematics project (without attempting to faithfully and completely represent the history and state of the art of these fields).

1) While mathematics has been carried out since antiquity, the systematic collection and uniform representation of mathematical knowledge took off only in the second half of the nineteenth century.

On a large scale, only a few attempts have been made over the last 150 years to collect all mathematical knowledge:

- F. Klein's (ed.) 23-volume, 20,000+ page *Enzyklopädie der mathematischen Wissenschaften* 1898 to 1933 [129],
- N. Bourbaki's (ed.) 9-volume *Elements of Mathematics* from 1934 to 2012 [126],
- M. Hazewinkel's (ed.) 13-volume *Encyclopedia of Mathematics* from 2002 [125] that derives from a project I. M.Vinogradov started in Russia in the 1970s, and
- P.R. Krishnaiah's and C.R. Rao's (eds.) (now) 32-volume *Handbook of Statistics* [149] which was started in 1984; volume 32 was just published earlier this year.

Selected smaller projects accumulating knowledge of expressly higher mathematics are:
- K. Ito's 2-volume *Encyclopedic Dictionary of Mathematics* (EDM2) from 1993 [127] and
- I. N. Bronshtein and K. A. Semendjajew's *Handbook of Mathematics* (first edition 1939) [125].

For a more complete listing of mathematical handbooks, see [128].

For selected areas of mathematics, especially when the mathematical knowledge is mostly ex-

pressed as formulas, various fairly comprehensive books have been published. Especially notable are Dickson's three-volume *History of the Theory of Numbers* from 1919 that lists number-theoretic results from the seventeenth century up to the book's publication. In a review of it from 1920, D. N. Lehmer said, "There is the greatest need for just such a piece of work to promote efficiency among the professional workers in this field and to prevent them from wasting their time on problems that have already been adequately treated, and also to suggest other problems which still defy analysis." It is this spirit of preserving, and making easily accessible, previously obtained mathematical knowledge that is today, nearly 100 years later, needed more than ever due to the exponential increase in mathematical knowledge. It can only be met in a truly digital, semantic, and computer-readable digital library of mathematics.

Tables of integrals remain among the most widely distributed mathematical books ever written [188]. Gradshteyn and Ryzhik (first edition 1943), Gröbner and Hofreiter (first edition 1944), and Prudnikov, Brychkov and Marichev's three-volume work are outstanding examples of accumulated, condensed, and unified mathematical knowledge. Handbooks about special functions, starting with Jahnke/Emde [133] from 1909, through the famous five volumes resulting from the Bateman Manuscript Project, to the legendary 1964 Abramowitz and Stegun [137] (which has nearly 70,000 citations according to Google Scholar) have proven to be invaluable collections of mathematical knowledge for mathematics itself and for the applied sciences.

We have not mentioned the various efforts to provide a comprehensive bibliography of mathematics that may be said to have culminated in the mathematics bibliography of Georg Valentin, tragically destroyed in Berlin during World War II. The modern successors of those efforts are the databases of Zentralblatt and Mathematical Reviews, which, while they do not explicitly organize mathematical theorems, represent the most comprehensive listings of where mathematical

knowledge is to be found. They should be of use in planning the projects of applying semantic markup to the corpus of mathematical literature.

2) Naturally, with the advent of the internet, new, general mathematical encyclopedias have arisen in purely digital form. Notable sites include:

- *MathWorld* [142]
- PlanetMath [143]
- Wikipedia [144]
- Scholarpedia [145]

Similar to specialized book-series projects, more specialized websites with mathematical content have been created over the last two decades. Examples include:

- Lie Atlas Project [136]: a website of representations of reductive Lie groups over real and *p*-adic fields
- Complexity Zoo [147]: a website with definitions for hundreds of computational complexity classes
- Statistics database [153]: a website with definitions for statistical distributions
- Polytopes [170]: a searchable database for polytopes
- Number Fields [171]: a searchable database for number fields
- LMFDB [60]: a database of *L*-functions and modular forms
- La Jolla Covering Repository: a database of good covering designs

As with many things on the web, despite being good and valuable projects, due to funding cycles, the fact that grad students and post-docs move on, and so on, not all such projects are well maintained. Notable examples of this phenomenon are the Encyclopedia of Combinatorial Structures [189] (last modified in 2011), the Code tables [191] (last updated in 2008), and the Design database (last updated in 2009).

3) In the 1990s, the first specialized digital collections of mathematical results that were (at least partially) designed for potential machine processing and use appeared on the internet. Not surprisingly, domains heavy on numerical and formula-based knowledge were among the first to be tackled. Large-scale, semantically rich mathematical libraries include:

- The On-Line Encyclopedia of Integer Sequences started by N. J. A Sloane, currently with about 255,000 sequences [138];
- The Wolfram Functions Site with 320,000 identities and about 10,000 visualizations [139] (this was the first larger fully human- and computer-readable and computationally ready digital mathematics library); and
- NIST's Digital Library of Mathematical Functions [140] with about 35,000 identities.

In addition to these websites, in recent years an orthogonal direction of semantic enrichment appeared—searching for mathematical formulas (the following is not an exhaustive listing):

- The Wolfram|Alpha [5] website allows for truly semantic search of mathematical formulas, in the sense that the question as well as the result are semantically fully understood (e.g., formulas can be dynamically generated, conditions and variable names are taken into account). Example queries include:

```
integral representations for zeta(y) valid for Re(y)>0
functional equation for sn(z,m)
DE for sin(x)*J(n,x)
```

This functionality is now also available in directly computable form via the "MathematicalFunction" entity type as part of extensive entity-property framework included in the recently released Version 10.2 of *Mathematica*.

- Literal search for formulas and formula fragments has been available for a number of years on various websites, e.g. the DLMF [104] (http://dlmf.nist.gov/help/search), the EuDML [2] (https://eudml.org/search) and on parts of the Springer website (http://latexsearch.com/), and Zentralblatt [10] (https://zbmath.org/formulae/).

4) Promising websites that contain quite technical information growing daily in quantity and powered by online communities [182] are:

- The nLab [141]—for a categorical view on modern mathematics
- MathOverflow [146]—a question and answer site for professional mathematicians and advanced students

A variety of other types of websites that accumulate mathematical knowledge by crowdsourcing

exists. For instance, the Wolfram Demonstrations Project [148] is a teaching-oriented one that offers thousands of interactive modules for a large spectrum of mathematical topics contributed by hundreds of mathematicians worldwide.

5) Symbolic computation software [154] started to appear in the 1960s. Today, the use of symbolic computation software in pure and applied mathematics is ubiquitous. In particular, a very large number of mathematicians now use programs such as *Mathematica*, Maple, and Magma on a daily basis in their research.

With the exception of the above listed websites on integer sequences, mathematical functions, and integrals and sums, most higher mathematics websites that are digital are not computable. Since the mid-2000s, Wolfram Research has pioneered the inclusion of mathematical data [152] into its *Mathematica* program, meaning that it includes collected, curated mathematical knowledge about various domains, e.g. graphs, lattices, finite groups, polyhedra and more. To a perhaps lesser extent, such knowledge is also implicitly encoded into other symbolic computation tools.

6) Closely related, but at the same time distinctly different, are the theorem provers and proof assistants written as standalone special programs. (One of the newest proof systems, *Theorema*, is a notable exception—it is built on top of *Mathematica*.)

While there are today specialized handbooks of certain mathematical fields that are quite dense in definitions, lemmas, and theorems (e.g., E. Pap's *Handbook of Measure Theory* [140]), "pure" collections of mathematical theorems with uniform notation and syntax are quite rare. The most notable collection is the 50,000+ theorems from pure mathematics [156] within the Mizar Mathematical Library [158] that have been assembled since 1970 in the dedicated journal *Formalized Mathematics* [157]. Although the emphasis of the project is on the proofs, and not on the

theorems, the large size of the theorem collection allows nontrivial research on the structure of mathematical theorems, such as investigating the distribution of the sentence complexity of these 50,000 theorems [159].

While not a collection of theorems, a collection of more than a million lemmas was recently assembled [194], [208] from proofs of theorems, which were then used for proving further theorems automatically. While the vast majority of these lemmas are not useful for direct human consumption, this project highlights the usefulness of large collections of machine-readable mathematical facts.

For some parts of mathematics, although not with the goal of collecting theorems but rather with the goal of proving them, collections of mathematical theorems for some specialized fields have been assembled. Examples are formalizations in the following areas:

- power series (see [178])
- real analysis ([181], see [87] for a comparison)
- algebraic numbers (see [77])]
- graph theory (see [113])
- complex plane geometry (see [81])
- constructive algebra (see [83], [202])
- category theory (see [84])
- Euclidean space (see [67])
- relational data models (see [178])

Another relevant aspect of proof systems is the experience their developers have gathered on the relation between mathematical structures and types, e.g. in [92], [114], [102], [46], [99], [52]. This aspect of semantic representations is intimately connected with point 8) discussed below.

7) The dominant classification of mathematics is the current 2010 Mathematics Subject Classification [160]. With 6,000+ classes it is the largest agreed-on subdivision of mathematics. Recently, the so-called *OntoMath*^{PRO} Ontology [23] was announced. It currently contains about 6,500 classes.

While the MSC 2010 and *OntoMath* are extremely useful classification schemes at the level of a paper, they do not uniquely characterize individual mathematical concepts. (Nor are they intended to.) Proper semantic encoding of mathematics clearly requires substantially more individual concepts than the ~6000 categories in MSC or *OntoMath*. A lower bound on the number of concepts is provided by the English Wikipedia, which contains roughly 15,000 pages devoted to specialized mathematical phrases [168]. However, the coverage on these pages is obviously far from comprehensive, and a conservative estimate of the number of distinct semantic objects needed to reasonably completely cover all major mathematical structures greatly exceeds 100,000 (cf. [199]). In particular, Wikipedia's list of named theorems is currently of length 1,100 by itself, whereas the actual number of extant mathematical theorems must be at least an order of magnitude larger.

For a tiny subset of relatively basic mathematics, the ISO 80000-2 standard [187] specifies the use of symbols and function names, but for the needs of research-level mathematics, this standard is not relevant. Unicode also lists over 3,000 special symbols that might be relevant to mathematical texts.

8) There are three aspects to a semantic representation language for mathematics: a) authoring and creating it, b) consuming and reading a mathematical statement written in it, and c) the mathematical coherence of the representation language itself.

As mathematicians are well aware, in TEX there is a substantial difference between the authoring form of a mathematical statement and its final display form. In recent years, a growing trend may be observed to advocate narrowing the gap between the authoring form and the final display form.

The consumption and reading aspect of a semantic representation of pure mathematics is closely

related to the use of mathematical notation. While mostly irrelevant for a computer, the use of "good" notations is critical for humans. The widespread use of TEX dramatically increased the quality and readability of mathematical papers and books. About 6,000 symbols are available with TEX to encode mathematical notations [162]. But even more important than the availability of a large number of special characters is the use of a notation that appears "natural" and historically in use in a certain field of mathematics. Only a very few books are dedicated to study of the notations of mathematics; the classic is F. Cajori's *History of Mathematical Notations* from 1928. The absence of a rich scholarly literature on this subject is also reflected in the MathOverflow community, where questions about mathematical notations are occasionally discussed.

While compact notations are useful for comprehension, they are often difficult for someone new to the field. A semantic markup language for mathematics should be as expressive as possible [193], [205]. A carefully designed language can dramatically help to speed up the understanding of theorems expressed in it [38]. A careful balance between being expressive, so that one comprehends a name for a function, operation or structure that one is not yet familiar with, and being concise has to be found in order to achieve a successful and usable semantic language for mathematics. (A good example of how expressive and powerful even small amounts of code can be,is the recent Tweet-a-Program initiative [166].) To achieve this goal, a good fraction of the $\sim 10^5$ mathematical concepts mentioned in point 7) will have probably to be defined in a semantic markup language for mathematics [15], [196].

The second aspect of the semantic markup language for mathematics is its authoring form. Only a uniform, expressive, rich and powerful language with a natural feel will be accepted by the mathematics community and stand the test of time. There is a large amount of computer science

literature on (computer) language design. A discussion of the general language of mathematics is in Ganesalingam's recent book *The Language of Mathematics* [98]; see also Muhammed's recent thesis [32] and Wiedijk's paper about the mathematical vernacular [177]. Various aspects of the unique challenges of designing a language for mathematics are discussed in [15], [165], [169], [172], [173], [204].

As the goal of the semantic markup language of mathematics is to express mathematics, rather than do a calculation, language design in this context does (mostly) not refer to computer language design (in the sense of programming paradigms, evaluation types, garbage collection). But because the markup language to be designed should also be understood by a computer, there are, of course, elements that are covered by the theory of computer languages that should be kept in mind. This is the third aspect, mentioned under c). For example, there are scoping rules for operations, operator precedences, and typing (in the sense of type theory). As B. Buchberger recently noted, "Today, however, mathematics and software are even more intimately connected by being the meta-theory for each other." Modern subjects of language design, such as co-data types and co-induction [174], [175], or algebraic effects [179] and more [195], are definitely of relevance for the semantic representation of mathematics.

Last, not least, and obviously (and so not listed as point 9), the semantic representation of mathematics should be based on the mathematics that it is designed to describe. While this is on the one hand an obvious statement, how concretely the mathematical content influences and shapes the representation is a matter to be discussed in detail in the workshop. Can the "spirit" of a mathematical field determine the representation? Or should it be the main theorems of a field? For example, P. Taylor argued "that the foundations for a mathematical discipline should be drawn from the headline theorems of that discipline itself."