Computable Data, Mathematics, and Digital Libraries in *Mathematica* and Wolfram|Alpha

Eric Weisstein

Wolfram|Alpha
Champaign, IL, 61820, USA
eww@wolfram.com

Abstract. This talk will focus on the infrastructure developed for representing and accessing data (especially mathematical data) in Wolfram|Alpha, as well as on the technologies and language extensions developed in the most recent version of *Mathematica* for making this data even more computationally accessible. Based on experiences using these technologies to create a prototype semantic digital library for a subset of mathematics, we believe the ambitious dream of creating of a semantic digital library for all of mathematics is now within reach.

1 Introduction

Wolfram|Alpha (`http://www.wolframalpha.com`) is a freely available website that contains hand-curated data sets taken from hundreds of technological, scientific, sociological, and other domains, including a core set of mathematical ones. This data has hitherto been accessible either directly via the website, through its API, or through a number of other specialized sources (such as various apps and SIRI). More recently, a large portion of this information has been exposed through the Wolfram Language itself via a set of built-in functions centered around an entity-property approach to information representation. The technology developed for Wolfram|Alpha has also recently been used and extended with the help of funding from the Sloan and Wolfram Foundations to create a prototype digital mathematics library covering known results and identities in the specific area of continued fractions. A recent US National Research Council report has identified approaches it believes could enable the creation of a substantial digital mathematics library, and we are currently investigating partnerships and technologies that could help turn this ambitious dream into a reality.

2 Computable Data in Wolfram|Alpha

Wolfram|Alpha was unveiled in 2009. In the ensuing years, it has become known for its ability to perform an extensive variety of computations in mathematics as well as many other fields. It currently answers millions of users queries per day. Whereas the traditional use of mathematical software is to carry out computations and the traditional use of encyclopedias is to give static information about

S.M. Watt et al. (Eds.): CICM 2014, LNAI 8543, pp. 26–29, 2014.

a certain entity or property, the goal of this website is to bring these two modes (purely dynamic and computational versus purely static and informational) together to dynamically generate knowledge about known structures.

Wolfram|Alpha's knowledge comes from a combination of *Mathematica* computations, roughly 1000 curated data sets, and links to a number of real-time data sources. Mathematical domains known to Wolfram|Alpha include graphs, groups, polyhedra, knots, curves, surfaces, and others. For querying computational knowledge, Wolfram|Alpha implements natural language encoding and processing. Finally, while results are by default displayed into a web browser, they are also available in a number of different formats including text, MathML, LaTeX, XML, images, together with *Mathematica* and its several data formats.

3 Computable Data in *Mathematica*

Starting in Version 6 (released in 2007), *Mathematica* itself has included a set of approximately 20 curated data collections covering mathematical, scientific, geographic, and a number of other domains. Five years after the release of Wolfram|Alpha and seven years after the released of *Mathematica* 6, the original concept of *Mathematica* data collections and the extensive additional functionality, coverage, and development work done for Wolfram|Alpha have been reunited in the recently released *Mathematica* 10.

There are several components to this integration, the first being the extensive augmentation of the set of available data collections. However, rather than bundling all the additional data into *Mathematica* itself, the integration has been accomplished using the Wolfram|Alpha API to expose a selected set of its data sets to *Mathematica* over an internet connection. One benefit of this approach is that data is updated, extended, and improved on the Wolfram|Alpha site much more frequently (usually weekly) than *Mathematica* itself is released.

An even greater step forward is the introduction of entity, entity class, property, and related built-in symbols as a means to represent and manipulate computable data in *Mathematica*. Each curated object in an available data set is assigned a domain (say `"PlaneCurve"`) and a canonical name (say `"Ellipse"`). Using this framework, objects can be easily referenced and acted open using functions such as `EntityValue[Entity["PlaneCurve","Ellipse"],EntityProperty["PlaneCurve","Area"]`, `EntityValue[Entity["PlaneCurve","Ellipse"],"Classes"]`, and so on. Similarly, a command like `EntityList[EntityClass["PlaneCurve","Conic"]]` can be used to list entities in the plane curve domain belonging to entity class conic.

There are also a number of convenient ways to construct or discover canonical entity, entity class, and property names from within *Mathematica*. The first is a revamped implementation of *Mathematica's* "free-form input" functionality. To wit, by preceding an input with a special character or keystroke (`=` for a simple Wolfram|Alpha result, `==` for a full result including all output pods, and `CTRL-=` for an in-line result), it becomes a natural language query to a Wolfram|Alpha server whose result is returned directly into the current notebook. For example,

simply typing `CTRL-= ellipse` into a *Mathematica* front-end returns the expression for the ellipse plane curve entity, while typing `CTRL-= ellipse area` gives the corresponding entity-property expression. As a trivial example, in the latter case, the resulting expression can be directly evaluated to give the expected formula `Function[{a, b}, π a b]`.

Not only does free-form input provide a simple interface for users to access data, it also provides a disambiguation mechanism in the event that multiple interpretations are available. For example, `CTRL-= mercury` defaults to a chemical element but presents the user with a set of assumptions for the plant, periodical, word, city, and given name. In a more computational setting, a programmatic approach is available using either `SemanticInterpretation` (which returns a single best semantic interpretation of the specified free-form string as a Wolfram Language expression) or `Interpreter` (which tries to interpret the natural language input as an object of the specified form).

The resulting synthesis of data representation, exposure, and access provides a powerful, flexible, and extensible framework which is practically applicable to virtually any domain of interest.

4 Prototype Semantic Digital Math Library: The eCF Project

Given the existence of the Wolfram|Alpha framework, it is natural to ask how difficult it would be to create from scratch a semantic digital library covering some specific domain of interest.

Precisely this question was addressed in the recently completed eCF ("e-Continued Fraction") project, undertaken from March 2012 to September 2013. The project resulted in the collection, semantic encoding, and exposure on the internet of significant results from the mathematical corpus concerning continued fractions. This work was supported by the Sloan Foundation with the goal of creating a new type of free digital archive for mathematical data that both ensures preservation and promotes dissemination of a targeted segment of mathematical knowledge for the public good.

Continued fractions presented an ideal subject for this proof-of-concept as they constitute a subset of mathematics that is historically rich, well-defined, and nontrivial, yet at the same time manageable in scope. Work completed includes a nearly exhaustive collection of continued fraction identities, a normalized representative bibliographic database of relevant books and articles, and an extensive collection of hand-curated theorems and results. All of these entities can be queried using a natural language syntax and provide additional linking and cross-entity entraining. In addition, many offer both visualizations and traditionally typeset versions, thus combining familiar traditional mathematical markup with modern tools for computational exploration.

This work was implemented using extensions of the framework developed for the Wolfram|Alpha computational knowledge engine and website. As such, it is generalizable to any area of knowledge where information is encodable and

computable. It differed from previous efforts by treating individual results (not papers) as entities of interest. Our methodology consisted of the following: 1) mine papers from archives of $\sim$ 800 historical articles, together with results from books and the newer literature, 2) extract theorems and other results, encode them in semantic form, and store them in computer-readable (and if possible, computational) form, 3) tag author, publication, reference, and subject information, 4) link to the original literature, 5) present in a coherent and unified form, 6) verify by human and computer, and 7) encode and access all data using extensions of the framework developed for Wolfram|Alpha.

At the completion of this work last year, approximately 400 theorems, conjectures, and other results were encoded and exposed. Results also include the first ever comprehensive table of continued fraction identities, containing $\sim$ 1,300 core and $\sim$ 11,000 derived continued fractions. All results are searchable using a natural language interface and are easily and freely accessible via Wolfram|Alpha.

5 Future Work

Our experience both with eCF and in other domains for which we have previously curated computable data in Wolfram|Alpha suggests it is feasible to develop tools and processes that allow a significant portion of mathematical knowledge to be mined, encoded, and exposed semi-automatically via crowdsourcing.

The Future World Heritage Digital Mathematics Library symposium took place in Washington, DC on June 1–3, 2012. After nearly two years of consideration, the National Research Council has now published their final report, which is available the arXiv e-print service (`http://arxiv.org/abs/1404.1905`).

While the NRC report is very detailed, touches on many aspects of relevance to the realization of a WMDHL, and in particular identifies approaches it believes could enable the creation of a substantial digital mathematics library, concrete steps that could be undertaken in the near-term to turn this dream into a reality remain elusive. A a result, the Sloan Foundation and Wolfram Foundation are currently investigating partnerships, technologies, and constituent components that could help turn the ambitious dream of creating a successful, comprehensive, and authoritative digital library for mathematics into a reality.

Acknowledgments. I thank the CICM organizers for the opportunity to share this work. I also express appreciation to Daniel Goroff, the Alfred P. Sloan Foundation, Stephen Wolfram, and the Wolfram Foundation for their support of the eCF project. I thank my eCF co-investigators Michael Trott, Oleg Marichev, Todd Rowland, and intern Christopher Stover. Finally, I thank Michael Trott and André Kuzniarek for helping spearhead the nascent effort to make the giant leap from "continued fractions" to "all of mathematics."