Can we find sequences without knowing any terms?: Bringing structured data to more of the OEIS

Eric Rowland

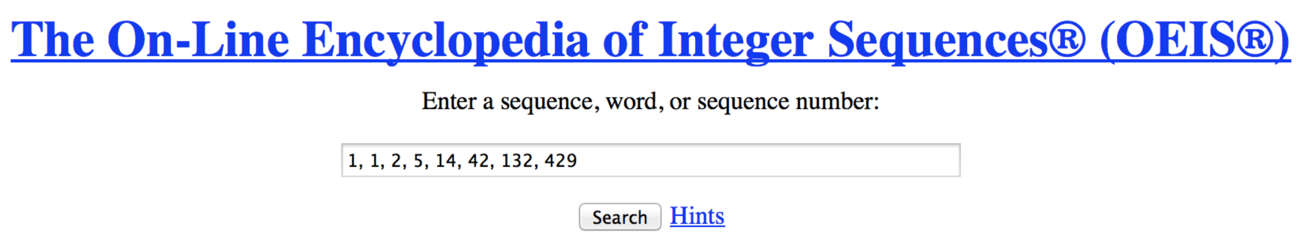
University of Liège, Belgium

Introduction

The OEIS [1] is an indispensable tool for identifying an integer sequence when one knows a few of its terms. However, there are potential applications of the OEIS that are currently infeasible because of the way data is collected and stored. By better structuring some of its data, the OEIS can become a useful tool for identifying and studying integer sequences in other ways as well.

Standard lookup

"What is this sequence $1, 1, 2, 5, 14, 42, 132, 429, \ldots$ that came up in my research?"

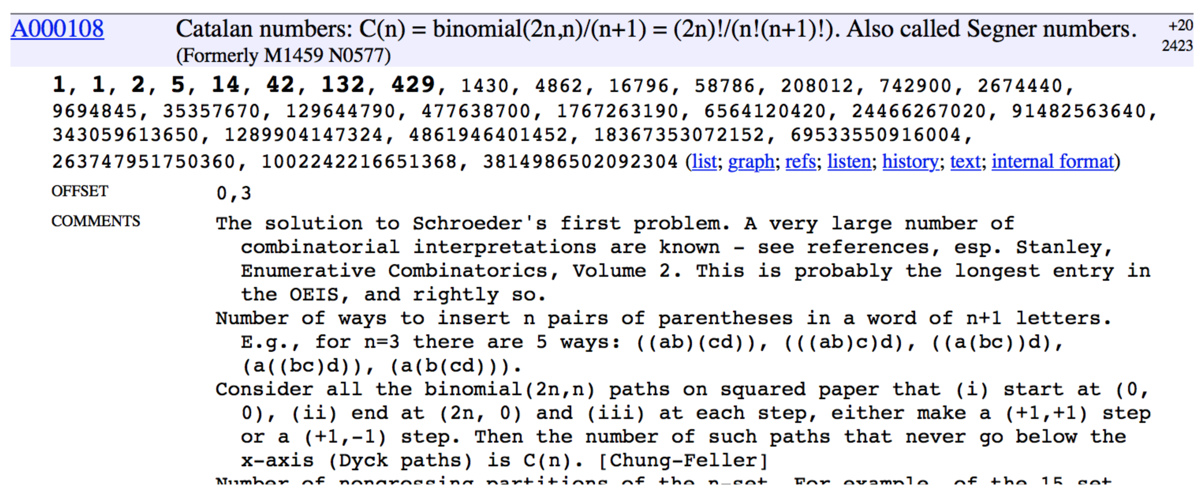

The On-Line Encyclopedia of Integer Sequences® (OEIS®)

Enter a sequence, word, or sequence number:

1, 1, 2, 5, 14, 42, 132, 429

Search Hints

It's the sequence of Catalan numbers! And the OEIS knows lots about it:


A000108 Catalan numbers: C(n) = binomial(2n,n)/(n+1) = (2n)!/(n!(n+1)!). Also called Segner numbers. +20 2423
(Formerly M1459 N0577)

1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, 2674440, 9694845, 35357670, 129644790, 477638700, 1767263190, 6564120420, 24466267020, 91482563640, 343059613650, 1289904147324, 4861946401452, 18367353072152, 69533550916004, 263747951750360, 1002242216651368, 3814986502092304 (list; graph; refs; listen; history; text; internal format)

OFFSET 0,3

COMMENTS The solution to Schroeder's first problem. A very large number of combinatorial interpretations are known - see references, esp. Stanley, Enumerative Combinatorics, Volume 2. This is probably the longest entry in the OEIS, and rightly so.
Number of ways to insert n pairs of parentheses in a word of n+1 letters. E.g., for n=3 there are 5 ways: ((ab)(cd)), (((ab)c)d), ((a(bc))d), (a((bc)d)), (a(b(cd))).
Consider all the binomial(2n,n) paths on squared paper that (i) start at (0, 0), (ii) end at (2n, 0) and (iii) at each step, either make a (+1,+1) step or a (+1,-1) step. Then the number of such paths that never go below the x-axis (Dyck paths) is C(n). [Chung-Feller]

In particular, the generating function is algebraic:

```
G.f.: A(x) = (1 - sqrt(1 - 4*x)) / (2*x). G.f. A(x) satisfies A = 1 + x*A^2.
```

Potential programmatic uses of OEIS entries

Suppose we have proved a new theorem about algebraic sequences (e.g. [2]). Now we want to find all algebraic sequences in the OEIS, to identify interesting examples and determine whether the new result solves any open problems. It seems the only way to do this is to extract entries with a line containing `G.f.`, parse equations on that line, and identify whether they are algebraic. This is complicated by the presence of multiple equations and minor inconsistencies (`A(x)` versus `A`).

Or suppose one wants to programmatically extract the Mathematica program from the **Mathematica** field of multiple entries. For the Catalan numbers, this is as follows.

```
MATHEMATICA   A000108[ n_ ] := (2 n)!/n!/(n+1)!
              A000108[n_] := Hypergeometric2F1[1 - n, -n, 2, 1] (* Richard L. Ollerton,
                Sep 13 2006 *)
              Table[ CatalanNumber@ n, {n, 0, 24}] (* Robert G. Wilson v, Feb 15 2011 *)
              CoefficientList[InverseSeries[Series[x/Sum[x^n, {n, 0, 31}], {x, 0, 31}]]/x,
                x] (* Mats Granvik, Nov 24 2013 *)
```

The first two lines define functions, whereas the last two lines output lists. Such inconsistencies are fine for human use but are an impediment to programmatic use.

Structured data

Data that is represented in a uniform way is immediately accessible to querying and manipulation by computer programs.

OEIS data is already structured into fields. Some fields contain data intended for humans, for example **Name**, **Comments**, and **Example**. Other fields contain precise structured data that is both human-readable and programmatically accessible; namely, the **Data** field contains the first few terms of the sequence as a comma-separated list, and this precise data format is what enables the standard lookup and Superseeker.

The **Formula** field and the fields containing code (in Maple, Mathematica, etc.) contain information that is potentially useful both to humans and to programs but is currently only accessible to humans.

To make this information usable by programs, we should represent it consistently and add any missing contextual information. For example, all elements of the **Mathematica** field should be of the same type, and this type should be as widely applicable as possible.

OEIS editors insure that conventions are respected in new entries and bring existing entries up to date, so implementation of new conventions can be crowd-sourced.

Tagging class information

There are many classes of integer sequences — periodic sequences, algebraic sequences, holonomic sequences, k-automatic sequences, sequences that count lattice paths, etc. One would like to have immediate access to sequences in each of these classes, along with standardized representations of them to facilitate computation of arbitrary terms, closure properties, and programmatic surveying and analysis.

A new **Classes** field could contain this information. For the Catalan numbers, this field would contain information such as the following.

```
AlgebraicGeneratingFunction:  x*A(x)^2 - A(x) + 1 = 0
HolonomicGeneratingFunction:  x*(4*x-1)*A''(x) + (10*x-2)*A'(x) + 2*A(x) = 0, A(0) = 1,
A'(0) = 1
HolonomicRecurrence:  (n+2)*a(n+1) - (4*n+2)*a(n) = 0, a(0) = 1
```

Each line contains the name of a class, along with a specification of the sequence in that class. The consistent notation allows programmatic interpretation.

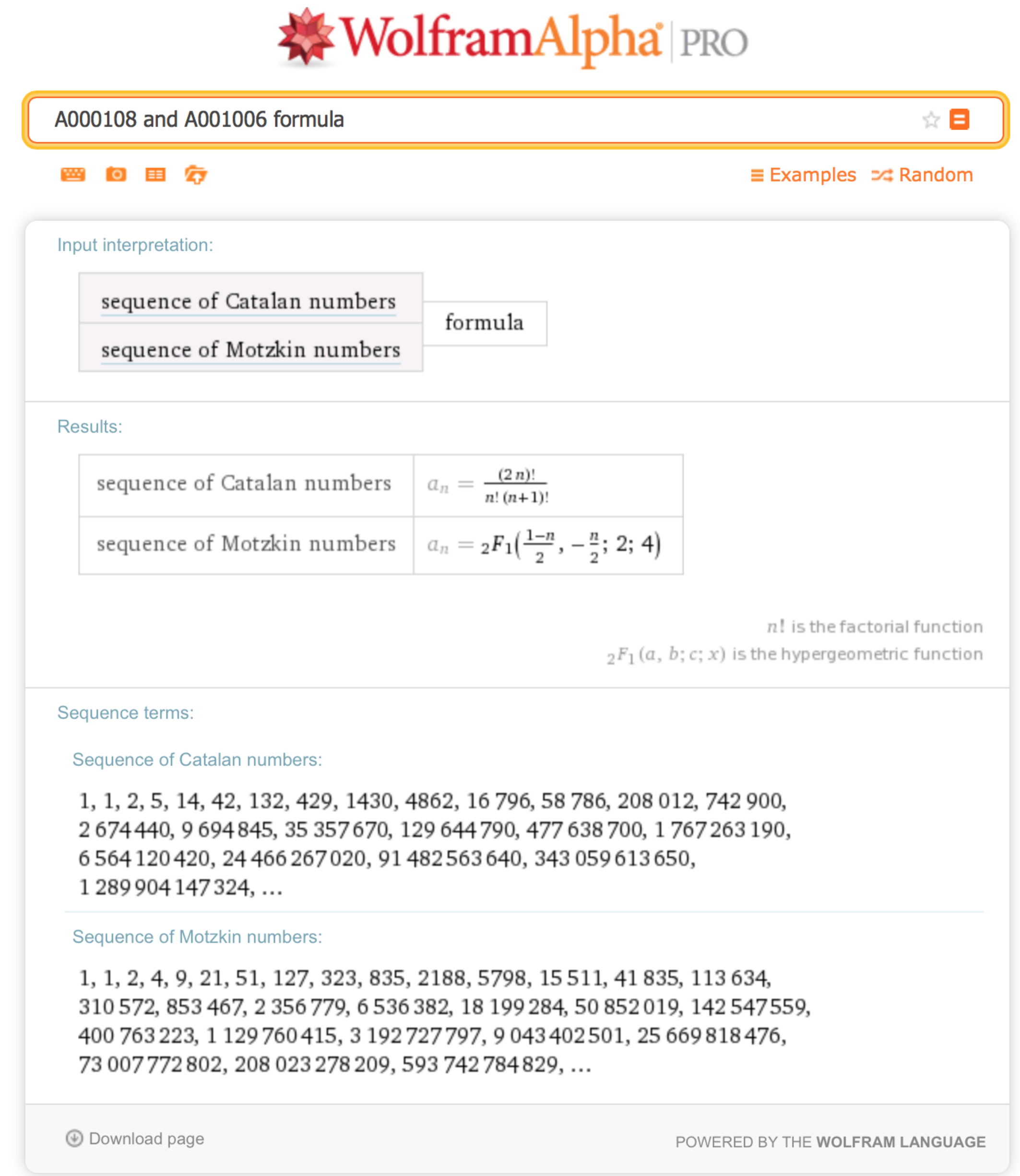
Benefits

By making OEIS data more computable, we will be able to...

- get statistical information about the relative number of sequences of various kinds that have arisen in human mathematics;
- search for sequences in new ways, allowing researchers to perform automated testing for examples or counterexamples to a given conjecture;
- do automated theorem discovery and proof by applying a general theorem to a large number of examples.

OEIS data in Wolfram|Alpha

Wolfram|Alpha [4] has incorporated OEIS data for a limited set of sequences, in a better structured form than the original.



Here we see programmatically-accessible formulas for the Catalan and Motzkin numbers, allowing computation of arbitrary terms.

Wikipedia's structured data initiative

A similar (much larger) initiative is currently underway to put Wikipedia data into structured form [3]. Like the OEIS, Wikipedia has amassed a large amount of precise data (dates, populations, geocoordinates, etc.). Wikidata is increasing the usefulness of this data by converting it into programmatically accessible form.

References

[1] The OEIS Foundation, The On-Line Encyclopedia of Integer Sequences, `http://oeis.org`.

[2] Eric Rowland and Reem Yassawi, Automatic congruences for diagonals of rational functions, to appear in *Journal de Théorie des Nombres de Bordeaux*, `http://arxiv.org/abs/1310.8635`.

[3] Wikidata: Introduction, `http://www.wikidata.org/wiki/Wikidata:Introduction`.

[4] Wolfram Research, Wolfram|Alpha, `http://www.wolframalpha.com`.