9 Data encodings

Key terms: `csv json pickle ETree sqlite3`

Reading: Python object serialization[19]

Exercise: Write a program that indicates how many of each element class appear in the XML file passed as command line argument.

You're familiar with zero-dimensional data, e.g. an `int`, and one-dimensional data structures, e.g. a `list` of `int`s. Perhaps you also worked with a two-dimensional structure, such as a comma-separated file (see the `csv` module), and learned not to mix up the dimensions. In general the process of writing multidimensional data to disk is called marshaling or serialization. Threre are many ways to go about it, and the affordances, efficiencies, and risks of the available file structures vary greatly. Here we will describe several of the principal schemes for serialization.

JavaScript Object Notation[20] is a structured file standard that offers a degree of human readability, and interoperability with other systems. The source of this notation is the JavaScript language, or alternately the family of ECMAScript languages to which it belongs. Its use tends to imply greater file length and the exclusion of custom structures. The `json` module's principal operations are `dump` and `load`.

`pickle` is specific to Python and more general: it can accommodate arbitrary Python objects. Since this includes arbitrary code, only load pickles that you wrote yourself, and don't expect others to load those you generate. Pickling provides a fairly fast and transparent means of moving data between memory and disk, which is often needed for large data sets. Files must be open for binary reading and writing, and can be automatically gzipped. Pickle formats are not permanent; they may change over time as new optimizations or developments are introduced. It replaces an earlier module called `marshal`.

`xml.etree.ElementTree`[21] is a minimal API for handling XML files. XML (eXtensible Markup Language) is a universal data encoding especially useful for document processing. It is a meta-language that encompasses formats such as HTML and SVG, and the more general case of XML databases. XPath is a simple and useful query language for XML content. XML documents can be vectors for attacks such as the "billion laughs" and the "quadratic blowup".[22] Many other tools[23] are available for specific use cases such as partial hierarchies and online processing.

```
# This tool enumerates the elements in any XML file.
# The implementation is a recursive depth-first search.

import sys, xml.etree.ElementTree as et, collections

def dive(depth,element):
 for child in element:
  print(depth*' '+child.tag)
  dive(depth+1,child)

with open(sys.argv[1]) as xml:
 content = et.fromstring(xml.read())
 dive(0,content)
```

More complex data storage options include `sqlite3`, a minimal implementation of the Structured Query Language for relational databases, and the host of databases with their own APIs.

Next week, we will study how features are determined and extracted from rich data.

[19] `https://docs.python.org/3/library/pickle.html`
[20] `https://docs.python.org/3/library/json.html`
[21] `https://docs.python.org/3/library/xml.etree.elementtree.html`
[22] `https://en.wikipedia.org/wiki/Billion_laughs_attack`
[23] `https://docs.python.org/3/library/xml.html`