The Open Dictionary Project

Tyler Nickerson
Linguistic Inc.
tyler@linguistic.io

Modern digital dictionary formats present numerous challenges for those seeking offline and programmatic access to lexical data. Most dictionary formats are built upon outdated or proprietary technologies, making them challenging to adopt in modern applications. This lack of standardization has led to a proliferation of custom protocols for storing data across many offline dictionary applications. This problem is further compounded by the fact that lexemes are typically stored as HTML fragments, making the extraction of structured data both difficult and variable across different formats.

In response to these challenges, we propose the Open Dictionary Project (ODict) as a solution for creating a new, open standard for dictionary formats that is both extensible and user-friendly. ODict includes a public file format specification, a command-line compiler, as well as wrapper libraries for creating and accessing lexical dictionaries. Dictionaries are compiled from human-readable XML and stored in a single, compressed binary file that can be easily transmitted and accessed on any device. Lexical information, such as etymologies, senses, and definitions, is stored in a buffer as structured data that can be output in a variety of formats. This approach allows language scientists to leverage well-formed lexical data in their own projects without having to process HTML.

Lexemes can be instantly retrieved *without* deserializing the file buffer and *without* relying on external index files, making lookups extremely fast. Furthermore, unlike other formats, ODict features official language bindings for NodeJS, Go, Python, and JVM languages (although the ODict specification can be used in any programming language supported by its underlying FlatBuffer protocol).

In this presentation, the process for writing, compiling, and consuming ODict files will be demonstrated, along with a discussion around the motivation and development of the format as described above. We will also share our future plans for the Open Dictionary Project in order to encourage a constructive discourse among conference attendees. The development of a new, open standard for dictionary formats is an important step towards improving the accessibility and reliability of lexical data for all users. We believe ODict has great potential in the fields of natural language processing and language learning, and we hope to see the format adopted by the wider community.