Dictionaries Conversion Framework

Oxford Global Languages Initiative by Oxford Dictionaries

In 2014 Oxford Dictionaries, part of Oxford University Press (OUP), launched its Oxford Global Languages initiative to empower millions of people across the globe with digital lexical data in 100 of the world's languages. As a part of its mission to extend learning and education worldwide, Oxford Dictionaries intends to make these languages available in digital form for free to everyone.

The challenge: efficient conversion process

Existing in-house dictionary data conversion required increased standardization and technology unification for input and output formats, as well as better quality control and assurance. The process was dependent on a set of unique specialists across the world and custom tools created by this team. Due to the manual efforts applied to conversion process, it could result in data loss and long lead time and excessive duration of the conversion process. Not to mention high costs of operations in this mode.

The brief

• Create a unified, automated and scalable data conversion process;

• Provide a one-box technology adjustable to any language and input/output formats;

• Reduce time and cost for the dictionary conversion process;

• Set up standards for quality control and assurance.

Oxford Dictionaries API
Enhance your app with our world-renowned dictionary data.
GET STARTED

The challenge was to turn dictionary conversion into an automated, time and cost efficient operation by creating a unified conveyor-based process adjustable to any language and data input/output formats.

Digiteum's role

It took Digiteum a month to analyze the dictionary conversion technology and the issues associated with the processing of large quantity of lexical content. We came up with a solution that could optimize the ongoing process and turn it into an efficient workflow.

First of all, Digiteum optimized the operations at the stage of data input and processing. Today the technology solves the issues associated with both ununified and unstructured linguistic data entry, the differences in language structures, and any errors at the input stage. As a result, the number of source formats increased from limited XML to an unlimited variation (XML, RTF, HTML, Plain Text, etc.), with data processing accuracy reaching desirable 100% in comparison with initial 15-20% data loss rate.

Secondly, Digiteum restructured the operations into an automated and scalable system. Previously the conversion was performed by a unique specialist and required an individual script for every single language. It was turned into a standardized scalable workflow performed by a role-based team of 2-6 interchangeable members. This optimization speeded up the process by about 10x and enabled simultaneous conversion of a number of dictionaries.

Digiteum's role

- Dictionary data conversion from 7 source formats into 3 target formats for any language;

- 2-5 days for one dictionary conversion (compared to up to 3 months of initial conversion terms);

- Scalable and automated conversion pipeline adjusted to any language and format and executed by a small team at significant costs savings;

- 100% data accuracy vs. 15-20% data loss.

Digiteum's role

Because of the efforts of the Digiteum team and OUP, OUP was able to complete phase one of the Oxford Global Languages (OGL) initiative. OGL is a major initiative from Oxford Dictionaries which aims to build resources for 100 of the world's languages and make them available online. In addition, the new conversion pipeline is also used for data for OUP data licensees and the OUP dictionary developer program. Since the new Dictionaries Conversion Framework was put into place more than 30 dictionaries in 18 languages have become available via the above programs.

The result

· OUP is equipped with a system of tools, processes, patterns, and .Net libraries for data conversion, which all together form Dictionaries Conversion Framework, that puts the Oxford Global Languages initiative in action;

· OUP scales up the coverage of pre-digital and digitally underrepresented languages;

· OUP uses Dictionaries Conversion Framework to effectively process and structure large quantity of quality lexical data;

· OUP performs analysis and processing of multilingual information, converts it into a decipherable format and provides easy access to this information at a global scale.

Find out more

For more information visit us online at:
www.digiteum.com/portfolio/oup-dictionaries

"Digiteum have proved to be a great partner for Oxford Dictionaries, with a strong and eclectic range of skills. They are incredibly resourceful in their approach to finding solutions and prepared to step in and take on anything from small to large initiatives. Thanks to Digiteum, we have made great strides on many of our projects, including our API platform, data conversion pipeline, and our websites." Judy Pearsall, Dictionaries Director, Oxford University Press.