Data-Driven Documentation:

A Technique for Reliable Multilingual Information Access

by Aarne Ranta,
University of Gothenburg and Digital Grammars AB, Gothenburg, Sweden.

In the globalizing world, both individuals and organizations have an increasing need to exchange information in different languages. For instance, a customer in France may want to order products from a Chinese company. The communication may involve initial browsing of Chinese web pages, followed by an order in English and a contract in French and Chinese. Each step involves information access in different languages, and the information itself should be exactly the same independently of language.

The traditional way of communicating between languages is by translation. In recent years, machine translation services such as Google and Baidu have become good enough for usage by consumers at browsing level. However, these techniques are not reliable enough for mission-critical tasks such as technical specifications and contracts. Such tasks are carried out by human translators, which is both slow and expensive.

In the talk, we will present an alternative technique for multilingual information access. The idea is to use machine-readable, formalized data as the ultimate representation of information. From this data, documents in different languages can be generated at need and with guaranteed fidelity. It is also possible to query the data in different languages and get answers in the same languages.

The technique of Data-Driven Documentation (DDD) has similarities with techniques such as natural language generation (NLG) , Controlled Natural Languages (CNL), and the Semantic Web. The novelty is in the integration of the different aspects (translation, querying, data acquisition) and, in particular, in the scalability and productivity of the approach. DDD is based on two decades of research on Grammatical Framework (GF), which has created resources and software covering 30 languages and enabling applications ranging from cloud services to mobile speech interfaces. The technique is currently being commercialized by the Digital Grammars AB company, but the GF technology itself is open source and freely usable by other parties as well.