# Verificação formal de uma implementação eficiente de UTF-8

**Leonardo Santiago**
leonardors@dcc.ufrj.br
UFRJ

## Abstract

O sistema de codificação *Unicode* é imprescindível para a comunicação global, permitindo que inúmeras linguagens utilizem a mesma representação para transmitir todas os caracteres, eliminando a necessidade de conversão. Três formatos para serializar *codepoints* em bytes existem, UTF-8, UTF-16 e UTF-32; entretanto, o formato mais ubíquito é UTF-8, pela sua retro compatibilidade com ASCII, e a capacidade de economizar bytes. Apesar disso, vários problemas aparecem ao implementar um programa codificador e decodificador de UTF-8 semânticamente correto, e inúmeras vulnerabilidades estão associadas a isso. Neste trabalho será utilizada verificação formal através de tipos dependentes, não apenas para enumerar todas as propriedades dadas na especificação do UTF-8, mas principalmente para provar que implementações estão conforme todas essas. Primeiro, uma implementação simplificada será desenvolvida, focando em provar todas as propriedades, depois uma implementação focada em eficiência e performance será dada, junto com provas de que as duas são equivalentes, e por fim essa implementação será extraída para um programa executável.

# Contents

# 1 Introdução