

BÀI 3: MỘT SỐ KIỂU DỮ LIỆU VÀ DỮ LIỆU VĂN BẢN

1. Phân loại và biểu diễn thông tin trong máy tính

Mã hoá thông tin thành dữ liệu nhị phân (dãy bit) được gọi là biểu diễn thông tin. Việc phân loại dữ liệu nhằm có cách biểu diễn phù hợp, tạo thuận lợi cho việc xử lý thông tin trong máy tính. Các kiểu dữ liệu thường gặp gồm: văn bản, số, lôgic, đa phương tiện (âm thanh, hình ảnh).

2. Biểu diễn dữ liệu văn bản

Cách mã hoá các kí tự được quy định trong bảng kí tự (bảng mã).

a) Bảng mã ASCII

ASCII là bảng mã chuẩn của Mỹ dùng để trao đổi thông tin. Bảng mã ASCII gốc sử dụng 7 bit, biểu diễn được 128 kí tự (chủ yếu là bảng chữ cái tiếng Anh). Bảng mã ASCII mở rộng sử dụng 8 bit, biểu diễn được 256 kí tự để bổ sung kí tự riêng cho các quốc gia khác.

b) Bảng mã Unicode và tiếng Việt trong Unicode

Unicode là bộ tiêu chuẩn biểu diễn kí tự văn bản hợp nhất quốc tế, sử dụng nhiều hơn 8 bit để mã hoá hàng trăm nghìn kí tự của hầu hết các ngôn ngữ trên thế giới. Chuẩn UTF-8 cho phép mã hoá kí tự với độ dài thay đổi từ 1 đến 4 byte. Đối với tiếng Việt theo chuẩn UTF-8: dùng 1 byte cho các kí tự La-tinh không dấu, 2 byte cho các nguyên âm có dấu cùng kí tự đ, Đ, và dùng 3 byte cho một số rất ít kí tự đặc biệt.

c) Số hoá văn bản

Tệp văn bản là định dạng lưu trữ ở bộ nhớ ngoài. Việc số hoá văn bản được thực hiện qua các phần mềm soạn thảo (như Word, Writer) hoặc thông qua công nghệ nhận dạng tiếng nói thành văn bản.