BÀI 3

MỘT SỐ KIỂU DỮ LIỆU VÀ DỮ LIỆU VĂN BẢN

SAU BÀI NÀY EM SẼ:

- Nêu được các loại thông tin và các kiểu dữ liệu sẽ gặp trong chương trình tin học phổ thông.
- Biết được các bảng mã thông dụng ASCII và Unicode.
- Giải thích được sơ lược về việc số hoá văn bản.

Thông tin đưa vào bộ nhớ máy tính dưới dạng các dãy bit. Như vậy khi đưa vào máy tính, phải mã hoá thông tin thành dữ liệu nhị phân. Tuỳ theo bản chất của thông tin được mã hoá mà dữ liệu tương ứng có cách biểu diễn riêng, từ đó hình thành nên các kiểu dữ liệu khác nhau. Vậy trong máy tính có các kiểu dữ liệu nào?

1. PHÂN LOẠI VÀ BIỂU DIỄN THÔNG TIN TRONG MÁY TÍNH

Hoạt động 1 **Phân loại thông tin**

Hình 3.1 minh hoạ một thẻ căn cước công dân. Trên đó có những thông tin gì?

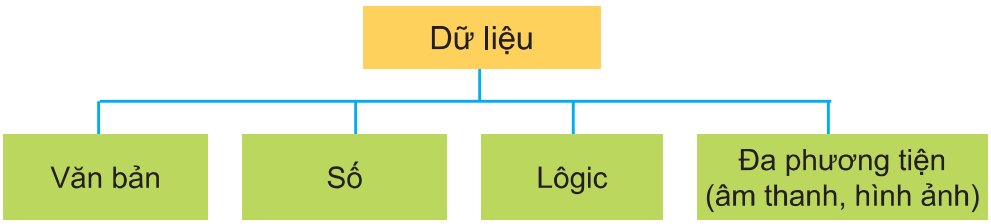
Hãy chia những thông tin đó thành các nhóm, ví dụ nhóm các thông tin có thể tách ghép được hay so sánh được để tìm kiếm và nhóm các thông tin có thể thực hiện được với các phép tính số học.

Hình 3.1. *Căn cước công dân*

Khi đưa vào máy tính thông tin được chuyển thành dữ liệu. Dữ liệu trên máy cũng cần được phân loại cho phù hợp với các phép xử lí trong máy tính. Ví dụ, đối với các dữ liệu là số có thể tính toán và so sánh. Còn đối với các dữ liệu dạng văn bản thì có thể tách, ghép, so sánh.

Việc mã hoá thông tin thành dữ liệu nhị phân được gọi là biểu diễn thông tin. Biểu diễn thông tin là bước đầu để có thể đưa thông tin vào máy tính.

Hình 3.2 là sơ đồ phân loại các kiểu dữ liệu được đề cập trong chương trình tin học phổ thông.



Hình 3.2. *Sơ đồ phân loại các kiểu dữ liệu*

- Biểu diễn thông tin trong máy tính là cách mã hoá thông tin.
- Các kiểu dữ liệu thường gặp là văn bản, số, hình ảnh, âm thanh và lôgic.
- Việc phân loại dữ liệu để có cách biểu diễn phù hợp nhằm tạo thuận lợi cho việc xử lí thông tin trong máy tính.

1. Theo em số căn cước công dân có kiểu số hay kiểu văn bản?
2. Kiểu số thực thường dùng để biểu diễn các số có phần thập phân (phần lẻ). Em hãy cho ví dụ một loại hồ sơ có dữ liệu kiểu số thực.

2. BIỂU DIỄN DỮ LIỆU VĂN BẢN

Việc đưa văn bản vào máy tính như thế nào không chỉ phụ thuộc vào kiểu dữ liệu là kí tự, xâu kí tự hay tệp văn bản mà còn phụ thuộc vào các kí tự ấy được mã hoá như thế nào? Cách mã hoá được quy định trong bảng kí tự.

Hoạt động 2 **Bảng chữ cái tiếng Anh và bảng chữ cái tiếng Việt**

1. Bảng chữ cái tiếng Anh có những kí tự nào?
2. Trong tin học, mỗi nguyên âm có dấu thanh của tiếng Việt là một kí tự. Hãy kể tên các kí tự tiếng Việt có trong bảng chữ cái tiếng Anh. Có bao nhiêu kí tự như vậy?

a) Bảng mã ASCII

Bảng mã được dùng phổ biến nhất trong tin học là "bảng mã chuẩn của Mĩ để trao đổi thông tin" (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange viết tắt là ASCII, đọc là as-ki). Ban đầu bảng mã này dùng các mã 7 bit, với 128 (2^7) mã khác nhau nên chỉ thể hiện được đúng 128 kí tự. Bảng mã 7 bit chỉ đủ dùng cho tiếng Anh, trong khi đó nhiều quốc gia có các kí tự riêng, như tiếng Hy Lạp có các kí tự α, β,γ; tiếng Nga có các kí tự б, г, д. Do đó, người ta đã mở rộng bảng mã 7 bit thành bảng mã 8 bit gọi là bảng mã ASCII mở rộng, cho phép mã hoá 256 kí tự, trong đó giữ nguyên 128 kí tự cũ. 128 vị trí được thêm vào trong bảng mã 8 bit so với bảng mã 7 bit được gọi là phần mở rộng của bảng mã ASCII. Các quốc gia có thể sử dụng phần mở rộng này cho các kí tự riêng của mình. Bảng mã ASCII với phần mở rộng chưa được thay thế bởi các kí tự riêng của các quốc gia được nêu trong Bảng phụ lục ở cuối sách. Trong bảng này, muốn lấy mã nhị phân của một kí tự thì chỉ cần ghép 4 bit ở chỉ số hàng với 4 bit ở chỉ số cột tương ứng với kí tự. Ví dụ mã nhị phân của "A" (có số thứ tự là 65) là 01000001.

b) Bảng mã Unicode và tiếng Việt trong Unicode

Ngoài các kí tự có trong bảng chữ cái tiếng Anh, tiếng Việt còn có 134 nguyên âm có dấu thanh và phụ âm "đ", kể cả chữ in hoa, đều không có sẵn trong bảng mã ASCII gốc, tuy nhiên phần mở rộng của bảng mã này lại chỉ có 128 vị trí.

Tình trạng thiếu vị trí còn trầm trọng hơn đối với các quốc gia dùng chữ tượng hình như Trung Quốc, Nhật Bản với vài chục nghìn kí tự. Chính vì thế, đầu những năm 1980, người ta đã đề xuất một chương trình quốc tế nhằm xây dựng một bảng mã hợp nhất, dùng chung cho mọi quốc gia, gọi là Unicode. Unicode thực tế là một bộ tiêu chuẩn biểu diễn kí tự văn bản trong máy tính, cho phép sử dụng nhiều hơn 8 bit để biểu diễn các kí tự thuộc nhiều ngôn ngữ khác nhau trên thế giới. Nhờ vậy, nếu bảng mã ASCII chỉ cho phép mã hoá 256 kí tự, thì Unicode hiện nay đã cho phép mã hoá hàng trăm nghìn kí tự khác nhau. Cùng với quy định, nếu quốc gia nào đã có một mặt chữ được xác định trước (bao gồm các kí tự gốc của bảng mã ASCII) thì quốc gia khác có thể dùng lại mà không phải định nghĩa một mã mới, Unicode tránh được tình trạng thiếu nhất quán do các quốc gia dùng các mặt chữ giống nhau nhưng mã khác nhau. Việc sử dụng Unicode tạo ra những ứng dụng đa ngôn ngữ, sử dụng đồng thời nhiều ngôn ngữ khác nhau như các trình duyệt web, ngôn ngữ lập trình, các phần mềm ứng dụng,...

Năm 2001 Việt Nam đã ban hành Tiêu chuẩn TCVN 6909:2001 về Bộ mã kí tự tiếng Việt 16-bit để sử dụng chung. Tiêu chuẩn này hoàn toàn phù hợp với tiêu chuẩn quốc tế về Unicode. Nó quy định mỗi kí tự đều được biểu diễn bằng 2 byte (các kí tự đã có trong bảng mã ASCII được bổ sung vào phía trước 8 bit giá trị 0). Năm 2017, Việt Nam cũng đã ban hành quy định bắt buộc sử dụng UTF-8 để biểu diễn bộ kí tự Unicode trong máy tính, trong đó sử dụng 1 byte để mã hoá các kí tự La tinh không dấu, sử dụng 2 byte để mã hoá các nguyên âm có dấu cùng các kí tự đ, Đ và chỉ dùng 3 byte để mã hoá một số rất ít các kí tự đặc biệt. UTF-8 (8-bit Unicode Transformation Format) là một trong các hệ thống định dạng chuyển đổi cho phép mã hoá kí tự với độ dài khác nhau (từ 1 tới 4 byte) dành cho Unicode.

Như vậy, hiểu một cách ngắn gọn, các bảng mã ASCII và Unicode quy định cách *biểu diễn kí tự*.

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 bằng các phần mềm soạn thảo văn bản như Word (của Microsoft) hay Writer (của Open Office). Gần đây người ta đã có thể nhập văn bản bằng nhận dạng tiếng nói. Chỉ cần đọc lời, máy tính có thể nhận dạng âm thanh và tạo ra văn bản.

- Bảng mã ASCII mở rộng sử dụng 8 bit để biểu diễn một kí tự.
- Unicode là bảng mã hợp nhất quốc tế, cho phép tạo ra các ứng dụng đa ngôn ngữ. Mỗi kí tự Unicode có thể được mã hoá bởi nhiều byte.

1. Sử dụng phụ lục Bảng mã ASCII mở rộng trang 165, hãy xác định mã nhị phân và mã thập phân của các kí tự S, G, K.

2. Trong bảng mã UNICODE, mỗi kí tự Tiếng Việt theo UTF-8 được biểu diễn bởi bao nhiêu byte?

A. 1 byte. B. 2 byte. C. 4 byte. D. từ1 đến 3 byte.

LUYỆN TẬP

1. Giấy phép lái xe có các thông tin nêu ở cột bên trái của bảng sau. Hãy ghép mỗi thông tin ở cột bên trái với kiểu dữ liệu thích hợp ở cột bên phải.

Thông tin	Kiểu dữ liệu
Ảnh	Số
Số	
Họ tên	Văn bản
Ngày sinh	Hình ảnh
Quốc tịch	Âm thanh
Nơi cư trú	

2. Câu trả lời nào đúng cho câu hỏi "Tại sao cần xây dựng bảng mã Unicode?"

A. Để đảm bảo bình đẳng cho mọi quốc gia trong ứng dụng tin học.

B. Bảng mã ASCII mã hoá mỗi kí tự bởi 1 byte. Giá thành thiết bị lưu trữ ngày càng rẻ nên không cần phải sử dụng các bộ kí tự mã hoá bởi 1 byte.

C. Dùng một bảng mã chung cho mọi quốc gia, giải quyết vấn đề thiếu vị trí cho bộ kí tự của một số quốc gia, đáp ứng nhu cầu dùng nhiều ngôn ngữ đồng thời trong cùng một ứng dụng.

D. Dùng cho các quốc gia sử dụng chữ tượng hình.

VẬN DỤNG

1. Dựa trên bảng mã ASCII, Việt Nam xây dựng bảng mã VSCII (Vietnamese Standard Code for Information Interchange), còn gọi là TCVN 5712:1993. Hãy tìm hiểu bảng mã này trên Internet theo những gợi ý sau:

– Bảng mã có đủ cho tất cả các kí tự tiếng Việt không?

– Bảng mã có bảo toàn bảng mã ASCII 7 bit không?

2. Phông chữ là hình ảnh của kí tự ứng với mã của kí tự. Không phải phông chữ nào cũng được thiết kế đầy đủ cho tiếng Việt. Hãy sử dụng phần mềm soạn thảo gõ một câu tiếng Việt và định dạng với các phông chữ khác nhau để tìm hiểu ngoài phông Times New Roman còn những phông nào đã thiết kế cho tiếng Việt Unicode.