## Учреждение образования

## Белорусский государственный технологический университет

Кафедра полиграфического оборудования и

системы обработки информации

**ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 3**

«Применение стандартов кодирования

текстовой информации»

по дисциплине «Стандартизация и сертификация

информационных систем и технологий»

Выполнил студент

ФИТ 4 курс 7 группа Бобрович Г.С.

Проверил

кандидат технических наук

Сулим П.Е.

Отчет по лабораторной работе

защищен с отметкой баллов

Минск 2024

***Цель* *работы*:**

Представление текстовой информации – разобраться со стандартами и ответить на вопросы.

Стандарты для следующих *coded character sets*:

* [*ASCII* *character set*](https://ru.wikipedia.org/wiki/ASCII)

ASCII — это таблица кодировки символов, в которой каждой букве, числу или знаку соответствует определенное число. В стандартной таблице ASCII 128 символов, пронумерованных от 0 до 127. В них входят латинские буквы, цифры, знаки препинания и управляющие символы.

С помощью ASCII вводят, выводят и передают информацию, поэтому она должна описывать самые часто используемые символы и управляющие элементы (перенос, шаг назад и так далее). Таблица восьмибитная, а числа, которые соответствуют символам, переводятся в двоичный код, чтобы компьютер мог их распознавать. Десятичное же написание удобнее для людей. Еще используют шестнадцатеричное — с его помощью легче представить набор в виде таблицы.

Заглавные и строчные буквы в ASCII — это разные элементы. Причем в таблице строчные буквы расположены под заглавными, в том же столбце, но в разных строчках. Так набор оказывается нагляднее, а информацию легче проверять и работать с ней, например редактировать регистр с помощью автоматических команд.

* [*ANSI character set*](https://ru.wikipedia.org/wiki/Американский_национальный_институт_стандартов)

Набор символов ANSI (ANSI character set) - множество 8-битных символов, включающее набор 7-битных ASCII-символов, а также символы валют, математические символы, подчеркнутые символы и другие символы, отсутствующие на клавиатуре.

Первая часть ANSI (entity numbers from 0-127) является оригинальным ASCII символов набора. Он содержит цифры, верхний и нижний регистр букв английского алфавита, а также некоторые специальные символы.

* [*Universal Coded Character Set*](https://ru.wikipedia.org/wiki/Универсальный_набор_символов)

Стандарт состоит из двух основных разделов: универсальный набор символов (UCS, Universal Character Set) и семейство кодировок (UTF, Unicode Transformation Format). Универсальный набор символов задаёт однозначное соответствие символов кодам — элементам кодового пространства, представляющим неотрицательные целые числа. Семейство кодировок определяет машинное представление последовательности кодов UCS.

Коды в стандарте Юникод разделены на несколько областей. Область с кодами от U+0000 до U+007F содержит символы набора Кириллица в Юникоде).

Универсальная система кодирования (Юникод) представляет собой набор графических символов и способ их кодирования для компьютерной обработки текстовых данных.

Графические символы — это символы, имеющие видимое изображение. Графическим символам противопоставляются управляющие символы и символы форматирования.

Графические символы включают в себя следующие группы:

* буквы, содержащиеся хотя бы в одном из обслуживаемых алфавитов;
* цифры;
* знаки пунктуации;
* специальные знаки (математические, технические, идеограммы и пр.);
* разделители.

[Юникод](https://ru.wikipedia.org/wiki/Юникод) — это система для линейного представления текста. Символы, имеющие дополнительные над- или подстрочные элементы, могут быть представлены в виде построенной по определённым правилам последовательности кодов (составной вариант, composite character) или в виде единого символа (монолитный вариант, precomposed character).

Освоить понятие *Unicode* кодеки: *UTF*-8, *UTF*-16, *UTF*-32.

[UTF-8](https://ru.wikipedia.org/wiki/UTF-8) использует от одного до четырех байтов на символ в зависимости от того, какой символ вы кодируете. Символы в диапазоне ASCII занимают только один байт, в то время как очень необычные символы занимают четыре.

[UTF-16](https://ru.wikipedia.org/wiki/UTF-16) использует два байта для большинства символов, четыре байта для необычных.

[UTF-32](ru.wikipedia.org/wiki/UTF-32) использует четыре байта на символ независимо от того, какой символ он есть, поэтому всегда будет использовать больше места, чем UTF-8 для кодирования одной строки. Единственное преимущество заключается в том, что вы можете рассчитать количество символов в строке UTF-32, только подсчитывая байты.

Освоить правила записей кодов символов.

Правила записи кода одного символа в UTF-8

1. Если размер символа в кодировке UTF-8 = 1 байт

Код имеет вид (0aaa aaaa), где «0» — просто ноль, остальные биты «a» — это код символа в кодировке ASCII;

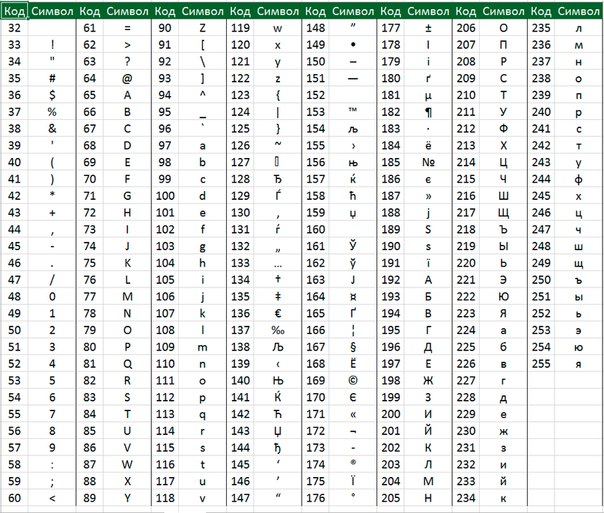
2. Если размер символа в кодировке в UTF-8 >1 байт (то есть от 2 до 6)

2.1 Первый байт содержит количество байт символа, закодированное в единичной системе счисления;

2.2 «0» — бит терминатор, означающий завершение кода размера

2.3 далее идут значащие байты кода, которые имеют вид (10xx xxxx), где «10» — биты признака продолжения, а «x» — значащие биты.

Привести примеры записей кодов символов



Разобраться в понятиях *code point*, *code unit*, *character set*, *coded character set*.

[Code point](https://en.wikipedia.org/wiki/Code_point): в терминологии кодировки символов, кодовая точка или позиция кода представляет собой любое числовое значение, составляющее пространство кода. Многие кодовые точки представляют собой отдельные символы, но они также могут иметь другие значения, например, для форматирования. Например, схема кодирования символов ASCII содержит 128 кодовых точек

[Code unit](https://en.wikipedia.org/wiki/Character_encoding#Terminology): единицы кодировки. Байт для utf-8, Слово (два байта) для unf-16 или Длинное слово (четыре байта) для utf-32.

[Character set](https://ru.wikipedia.org/wiki/Набор_символов): таблица, задающая кодировку конечного множества символов алфавита (обычно элементов текста: букв, цифр, знаков препинания).

[Coded character set](https://en.wikipedia.org/wiki/Universal_Coded_Character_Set): содержит привязку каждого символа из репертуара к целому неотрицательному числу, называемому кодовой точкой (code point);

Привести примеры кодирования текстовой информации.

Способы кодирования информации:

1. при помощи чисел — числовой.
2. Кодирование при помощи знаков того же алфавита, что и исходный текст — символьный.
3. Кодирование при помощи рисунков и значков — графический.

Приведите краткие выводы по содержанию работы.

Вывод: были разобраны различные виды символьной кодировки, их стандарты и способы кодирования текстовой информации. Разобраны базовые понятия по теме