## Учреждение образования

## Белорусский государственный технологический университет

Кафедра полиграфического оборудования и

системы обработки информации

**ЛАБОРАТОРНАЯ РАБОТА № 3**

по дисциплине «Стандартизация и сертификация

информационных систем и технологий»

Тема

«Применение стандартов кодирования

текстовой информации»

Выполнил студент

Шалунов Дмитрий Леонидович

Проверил

кандидат технических наук

Сулим П.Е.

Отчет по лабораторной работе

защищен с отметкой баллов

Минск 2023

***Цель* *работы*:**

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

Стандарты для следующих *coded character sets*:

* *ASCII* *character set*
* *ANSI character set*
* *Universal Coded Character Set*

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

Освоить правила записей кодов символов.

Привести примеры записей кодов символов

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

Привести примеры кодирования текстовой информации.

Вывод:

Форматы хранения данных: txt, docx, json, … .

Кодировка – процесс перевода данных из формы удобной для восприятие в форму удобную для хранения.

ASCII – однобайтовая кодировка, прородитель всех.

Windows 1251 – ASCII + национальные языки

ANSI – зависит от реестра.

Unicode (UTF-16, …) – многобайтовая кодировка с переменной длинной символа.

Code unit – основной компонент, используемый системой кодирования символов.

Code point – число, которое соответствует определенному символу.

Порядок кодирования в UTF-8:

* Каждый символ превращаем в Юникод.
* Проверяем из какого символ диапазона.
* Если код символа меньше 128, то к результату добавляем его в неизменном виде.
* Если код символа меньше 2048, то берем последние 6 бит и первые 5 бит кода символа. К первым 5 битам добавляем 0xC0 и получаем первый байт последовательности, а к последним 6 битам добавляем 0x80 и получаем второй байт. Конкатенируем и добавляем к результату.
* Похожим образом можем продолжить и для больших кодов, но если символ за пределами U+FFFF придется иметь дело с UTF-16 суррогатами.

Порядок декодоривоная UTF-8:

* Ищем первый символ вида 11xxxxxx
* Считаем все последующие байты вида 10xxxxxx
* Если последовательность из двух байт и первый байт вида 110xxxxx, то отсекаем приставки и складываем, умножив первый байт на 0x40.
* Аналогично для более длинных последовательностей.
* Заменяем всю последовательность на нужный символ Юникода.

