**BC/NW 2023 № 1 (40):11.2**

**МЕТОДЫ ГЕНЕРАЦИИ СИНТЕТИЧЕСКИХ ДАННЫХ ДЛЯ ОБЕЗЛИЧИВАНИЯ**

Бабак Н.Г., Оцоков Ш.А

Чтобы заменить персональные данные (ПДн) и прочую конфиденциальную информацию на обезличенные [1] значения синонимическим алгоритмом необходимо сгенерировать синтетические данные. Генерация должна осуществляться по определённым правилам, а также при синонимической замене сгенерированные обезличенные сущности должны быть неотличимы от исходных данных.

В зависимости от типа данных и их предполагаемого использования существует несколько основных способов их генерации:

1. Формирование синонимичной сущности с помощью генератора псевдослучайных чисел и расчёта контрольных разрядов;
2. Замена значением из словаря с сохранением морфологических [2] особенностей текста;
3. Формирование заглушки на основе типа заменяемой сущности;
4. Размытие числа или даты в заданном диапазоне.

К правилам, на основе которых генерируется синтетическое значение, относятся тип и вид данных, морфологические особенности (род, число, падеж, часть речи), наличие контрольного разряда, алгоритм обезличивания. Например, при использовании в качестве алгоритма обезличивания маскирования в большинстве случаев достаточно генерации нескольких заглушек/плейсхолдеров. В то время как для синонимического обезличивания используются практически все перечисленные способы генерации синтетических данных.

Наибольшую сложность в структурированной информации представляют поля, содержащие несколько подтипов данных. Например, поле с клиентами может содержать как физических лиц – фамилия, имя, отчество, так и юридических лиц – полное или краткое название организации. В таком случае при генерации необходимо учитывать вид данных в каждой отдельной ячейке.

Также при генерации данных необходимо учитывать их формат записи и занимаемую длину, особенно для неструктурированной информации. Кроме того, одинаковые значения могут быть записаны в разном формате, что тоже необходимо учитывать, так как одинаковые значения при обезличивании следует заменять на одинаковые синтетические.

**Литература**

1. **Приказ Роскомнадзора** от 5 сентября 2013 г. № 996 «Об утверждении требований и методов по обезличиванию персональных данных».
2. **Korobov M.** Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332 (2015).