Наталья Иванова
Максим Шашков
gbif.ru@yandex.ru

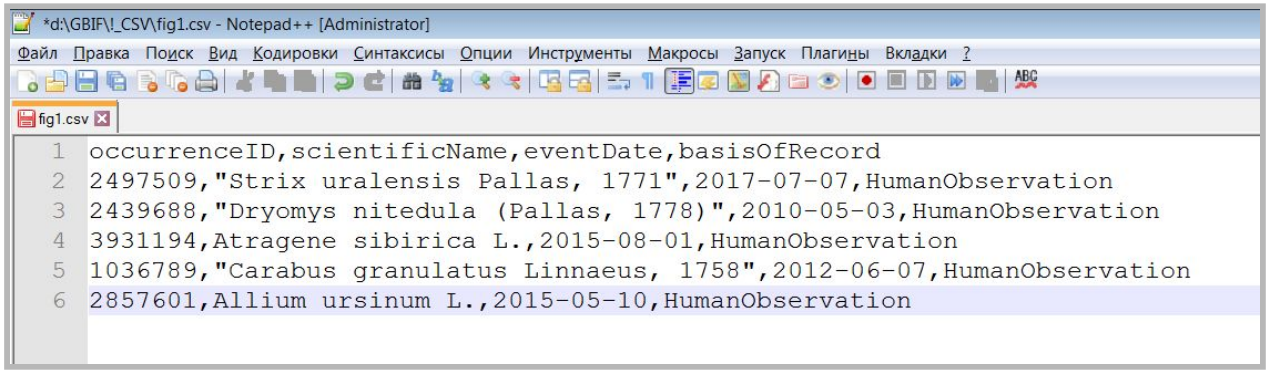
Сохраняйте CSV-файлы в гармонии с окружающим миром

Перед тем, как загрузить таблицу в IPT, желательно сохранить ее в формате CSV (от англ. Comma-Separated Values — значения, разделённые запятыми). Это универсальный формат, предназначенный для представления таблиц в виде простого текстового файла и служащий для обмена данными.

Для чего это нужно? В формате CSV данные хранятся “как есть”. EXCEL автоматически преобразует ваши данные при открытии: изменяет формат представления дат, округляет длинную дробную часть чисел, удаляет лидирующие нули (09 будет преобразовано в 9) и др. В результате такой "обработки" все ваши усилия по корректировке и стандартизации данных могут оказаться потраченными впустую. Поскольку CSV - это текстовый формат, файл будет занимать меньший объем, чем файл EXCEL и его обработка потребует гораздо меньше системных ресурсов: попытка загрузить файл EXCEL с несколькими десятками тысяч записей запросто может "повесить" IPT. Кроме того, CSV файл откроется на любом компьютере с любой операционной системой.

Как это выглядит? Как написано выше, CSV - это файл, где значения разделены запятыми. Это означает, что знак "запятая" является для компьютера указателем, по которому он определяет границы между столбцами таблицы. Если же запятая является частью какого-то значения, то такая ячейка таблицы заключается в кавычки.

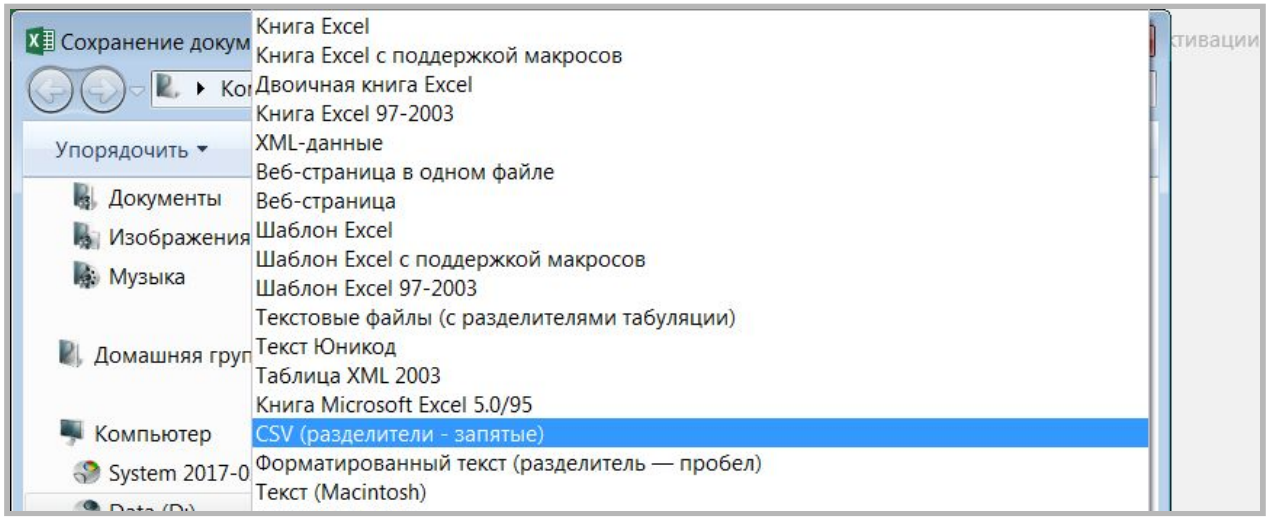
Вот так выглядит CSV файл в текстовом редакторе notepad++



Этот же файл, открытый в EXCEL

	A	B	C	D	E
1	occurrenceID	scientificName	eventDate	basisOfRecord	
2	2497509	Strix uralensis Pallas, 1771	2017-07-07	HumanObservation	
3	2439688	Dryomys nitedula (Pallas, 1778)	2010-05-03	HumanObservation	
4	3931194	Atragene sibirica L.	2015-08-01	HumanObservation	
5	1036789	Carabus granulatus Linnaeus, 1758	2012-06-07	HumanObservation	
6	2857601	Allium ursinum L.	2015-05-10	HumanObservation	
7					
8					
9					

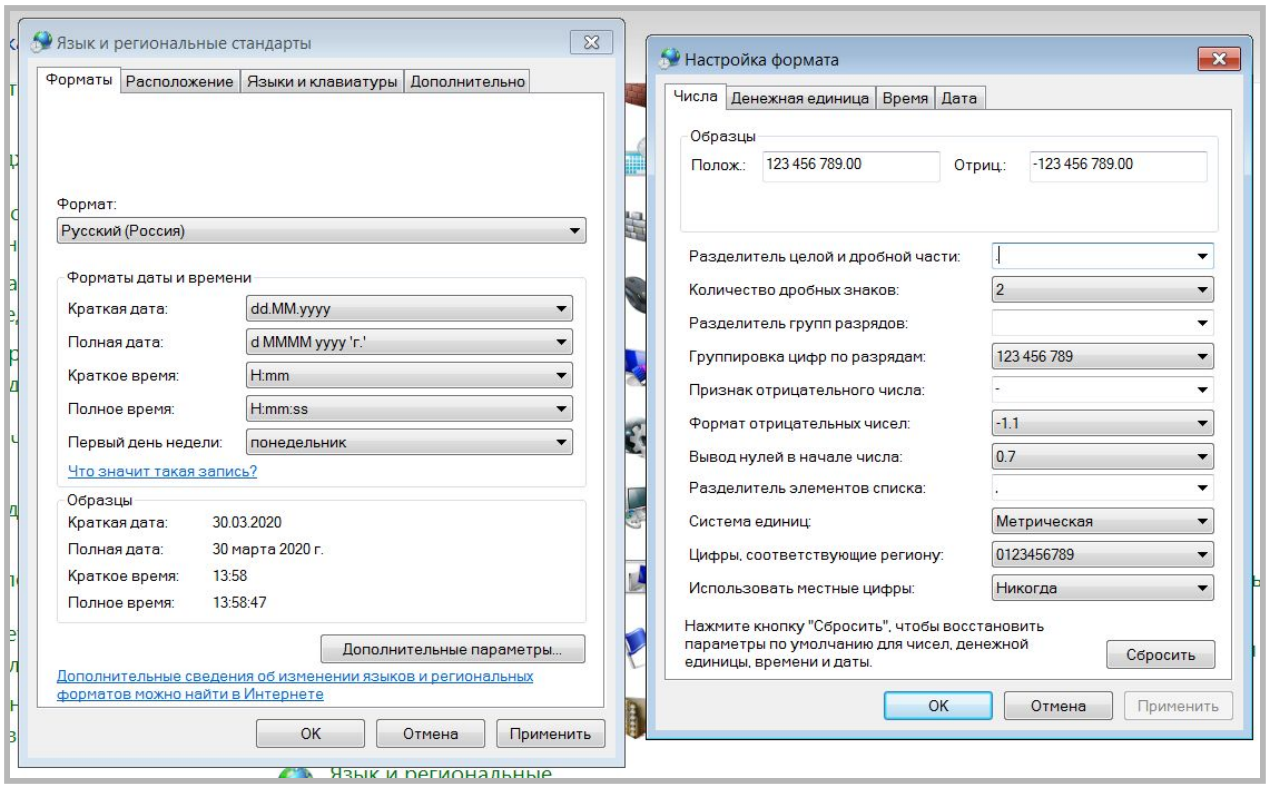
Как сохранить таблицу в CSV? Чтобы в программе EXCEL сохранить файл в формате CSV нужно воспользоваться функцией "Сохранить как", доступной в меню "Файл". **НО!** Вы обнаружите, что можно выбрать несколько вариантов сохранения CSV файла.



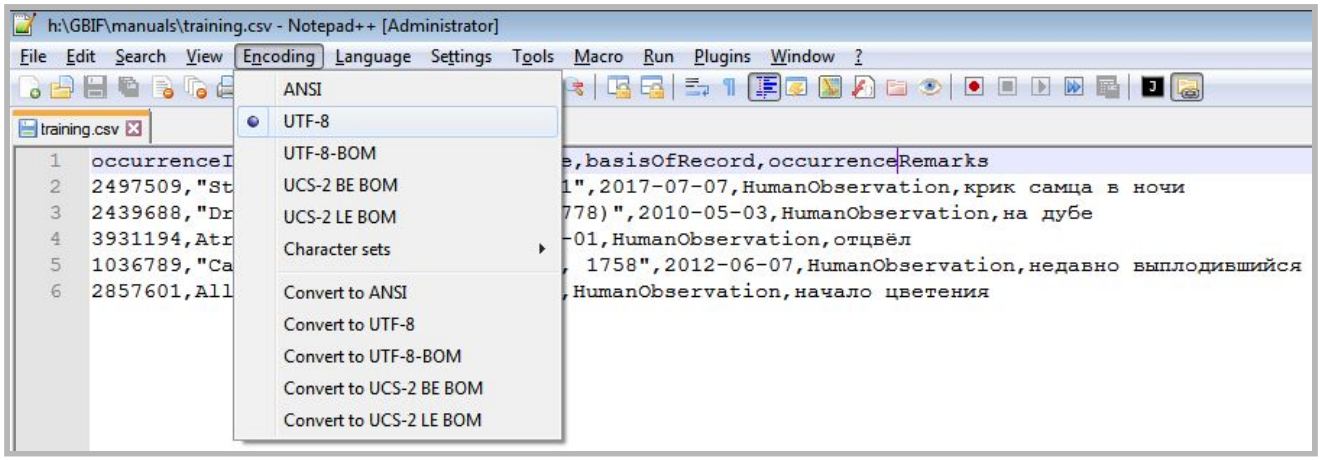
Причина в том, что в России запятая, как правило, используется для разделения целой и дробной части в числах (международный стандарт - точка). Поэтому в файлах CSV в качестве разделителя у нас часто используют знак табуляции или точку с запятой. В результате получаются файлы, которые, строго говоря, CSV уже не являются (см. статью в Википедии), а при загрузке их в IPT или пересылке коллегам возникает множество проблем.

Что делать? Сохранять файл в правильный CSV с разделителем полей "запятая".

Для этого нужно выбирать вариант "CSV (разделители - запятые)". В операционной системе нужно изменить разделители: указать разделитель целой и дробной части - "точка", разделитель полей (элементов) - "запятая". Эта опция доступна через меню Пуск -> Панель управления -> Язык и региональные стандарты -> Дополнительно.



После того, как CSV файл будет сохранён, нужно обязательно проверить кодировку символов, особенно, если в файле есть буквы кириллического алфавита или еще какого-то, отличного от латинского. Для русского языка могут использоваться 4 разные кодировки: MC866, Windows-1251, KOI-8 и UTF-8. Для отправки данных в IPT использовать надо последнюю. Файл можно проверить, открыв его в Notepad++, если кодировка не соответствует - конвертировать в UTF-8.



Обратите внимание, что в IPT существует техническая возможность загружать CSV файлы четырех вариантов, с разделителями в виде запятой (Comma), точки с запятой (Semicolon), табуляции (Tab) или вертикальной черты (Pipe). Для этого необходимо указать соответствующий вариант в поле "Разделитель полей (Field Delimiter)". Но чтобы меньше путаться в форматах, лучше взять себе за правило всегда сохранять данные в “правильный” CSV файл.