Импорт росписей из внешних форматов (плагин GEDKeeper)

# Общее описание программы и особенностей работы

Плагин «Импорт росписей из внешних форматов» предназначен для импорта росписей из таких файлов как текстовые (txt), документы Word (doc) или таблиц Excel (xls). Таблицы Excel поддерживаются ограниченно, без настройки входного формата.

Создать полностью интеллектуальную программу импорта из любых произвольных текстовых данных очень сложно, поэтому есть ряд правил и ограничений.

В общем случае, программа импорта рассчитывает на следующий формат данных:

Поколение I

1. Иванов Иван, \*1680 +1725

Текстовый комментарий.

Ж – Иванова Мария Петровна, \*1685 +1740

Текстовый комментарий жены.

Поколение II

1-1. Иванов Иван Иванович, \*1700 +1757

Тестовые данные, комментарии.

Т.е. роспись состоит из строк, каждая из которых может быть отнесена к одному из трех вариантов: строка заголовка поколения, строка персоны с датами рождения/смерти и текстовый комментарий (текст, описание, краткая биография).

Заголовок документа до любой строки поколения или строки персон игнорируется.

Строка заголовка поколения может иметь один из двух заложенных форматов: либо только римское число (I, II, III, IV, V и т.д.), либо «Поколение N», где N – может быть как римским, так и арабским числом. Другие форматы этой строки не будут распознаваться, но если требуется – необходимо обратится к автору программы с описанием необходимого формата.

***Внимание: строка заголовка поколения информационного значения для программы и базы данных не несет, поэтому может отсутствовать, т.е. не является обязательной.***

Программа разбивает входные данные на строки, после чего пробует найти в строке характерные признаки персональных данных. Если импортируется файл документа Word, то нужно понимать, что один абзац такого документа без разрывов – это одна длинная строка для программы (это нормально).

Строка персональных данных определяется по следующему критерию: она должна начинаться с порядкового номера персоны определенного формата. Форматы, которые заложены в программу: нумерация по д`Абовиллю и нумерация по Коновалову.

В случае нумерации персон по д`Абовиллю строки должны начинаются с: «1.», «1.1.», «1.1.1.», «1.2.1.» и т.д.

В случае нумерации персон по Коновалову строки должны начинаются с: «1.», «2-1.», «3-2.», «4-1.» и т.д.

***Внимание: программа в общем случае рассчитывает, что в строке персональных данных, кроме номера, полного имени человека, и дат рождения/смерти больше никаких данных нет. Все прочие данные должны указываться в последующих строках текста биографии.***

В процессе импорта, программа определяет каждую строку персональных данных, после чего информация из неё обрабатывается следующим образом:

1. Номер:
   1. Номер в первую очередь служит для проверки целостности данных, т.е. номера должны быть последовательными, без разрывов. При наличии разрывов программа будет считать данные неполными, с низким доверием к входной информации.
   2. Непрерывность номеров необходима для основной задачи: определения вертикальных родственных связей в генеалогических данных росписи. Т.е. без правильных номеров выстроить корректное древо «предок -> потомок» невозможно.
   3. В случае нумерации по Коновалову номер также *может* содержать порядковое число брака, в котором родился данный человек.
2. Полное имя человека
   1. Имя определяется как часть строки после номера и до классических обозначений дат рождения («\*») и смерти («+»). Если имя заканчивается на точку или запятую – они отбрасываются.
   2. *Внимание: программу трудно научить безошибочно определять формат имен «Фамилия Имя Отчество» или «Имя Отчество Фамилия», поэтому формат имён в росписи задается пользователем до начала основной стадии импорта.*
   3. *Внимание: сложно создать корректное распознавание всевозможных форматов исторических вариантов имен, поэтому если автор документа использовал, например, такой формат «Иван (Ивашка) Васильев (Васильевич) Петров» - такое имя будет занесено в базу данных программы с ошибками. Более того, зачастую, даже зная, что формат имен в росписи, например, «Имя Отчество Фамилия», если программа встречает имя «Ивашка Васильев» - трудно определить, «Васильев» относить к отчеству или фамилии? (здесь приведен очень простой пример, в реальных данных бывают случаи, когда вторым словом идет и отчество, но фамилии нет, или фамилия без отчества, или есть только имя и прозвище). Еще хуже, когда всё это вперемешку с частями в скобках.*

# Реализовано и работает

Распознается:

1. Форматы нумерации по д`Абовиллю и Коновалову.
2. В персональных строках определяются даты рождения и смерти, при наличии стандартной символики «\*» и «+», в т.ч. относительные (поддерживаются следующие варианты: «п.», «до», «ок.» и «около»).
3. Распознаются два основных формата персональных имен: «Фамилия Имя Отчество» и «Имя Отчество Фамилия».
4. Распознаются строки супругов следующих форматов:
   1. «Ж/М - ФИО»
   2. «ЖN/МN - ФИО», где N – номер брака;
   3. «ЖN/МN (дата брака) - ФИО», где N – номер брака;
   4. «Ж/М: ФИО»
   5. «ЖN/МN: ФИО», где N – номер брака;
   6. «ЖN/МN (дата брака): ФИО», где N – номер брака;
5. В строке супругов определяются даты рождения/смерти также как у основных персон росписи. В т.ч. относительные.
6. В строке супругов определяются даты брака и подставляются в соответствующий факт семейной записи.

# Текущие проблемы

1. Не распознаются даты, заданные в формате "+1710 от родимца" - т.е. текст после даты не обрабатывается отдельно. Основание: в различных росписях текст после даты может быть местом рождения/смерти, или причиной смерти, или одновременно и то и другое.
2. Если формат имени задан как "ИОФ", а в строке два компонента, один из которых фамилия (например, Иван КОЛЕСОВ), то фамилия учитывается как отчество. Как сделать правильное распознавание компонентов имен? Можно попробовать сделать распознавание фамилии по верхнему регистру (но так бывает не во всех документах). Либо делать интеллектуальный эвристический словарь имен/отчеств.
3. Текст комментариев к супругам остается в комментарии к основной персоне. Делать перенос комментария к супругу или нет?
4. Если в имени есть посторонние компоненты (возможные варианты в скобках и без) - они не будут отработаны.
5. В настоящее время импорт росписей по д`Абовиллю заблокирован, т.к. нуждается в комплексной отладке. Более того, во многих найденных европейских родословных ситуация усугубляется тем, что информацию по персоне смешивают в кучу в одной строке. Причем зачастую она идет вперемешку, т.е. в одной строке даты могут быть до указания супругов, в другой – после, в третьей ко всему этому еще добавляются титулы и места проживания/рождения/смерти.

# Как работать с программой

1. Для начала необходимо открыть файл импортируемой росписи в Word’е или Excel’е и запомнить следующее:
   1. Тип нумерации персон;
   2. Формат имен «ФИО»/«ИОФ»;
   3. Формат строк поколений (это только римские номера или другое);
   4. Особенность формата дат рождения и смерти в персональных строках (заключены они в скобки или нет).
2. В программе GEDKeeper создать новый пустой файл базы данных (это обязательно, т.к. процесс скорее всего будет в несколько этапов и нежелательно замусоривать базы данных с полезной информацией).
3. Открыть плагин «Импорт росписей из внешних форматов».
4. Кнопкой «Выбрать» открыть файл росписи (предварительно он должен быть закрыт в Word’е или Excel’е – если еще открыт).
5. Необходимо подождать, пока плагин выполнит предварительную загрузку и анализ данных.
6. Если это тестовый документ, то автоматическое определение формата номеров будет выполнено сразу. Необходимо только проверить, что программа определила этот формат верно (блок опций «Формат нумерации персон»). Т.к. формат Excel-файлов сейчас заложен только один, предварительный анализ формата не делается, там могут быть только номера по Коновалову.
7. Далее, необходимо проставить те опции, которые сейчас уже обрабатываются программой (они не закрашены серым цветом):
   1. Формат ФИО;
   2. Формат строки поколений;
   3. Если в исходном файле даты рождения и смерти персон заключены в скобки – нужно отметить специальную настройку №1.
   4. Если требуется «нормализовать» фамилии персон из верхнего регистра в нижний регистр с заглавной первой буквой – поставить соответствующую опцию в блоке «Параметры преобразований».
8. Нажать кнопку «Вперед >» и подождать пока программа выполнит анализ данных и их основной импорт в базу данных.
9. На этом шаге нужно внимательно проанализировать список ошибок импорта. Каждая ошибка выделяется знаками «>>>>».
10. Если есть ошибки и необходим как можно более чистый импорт с максимальным переносом данных в базу – нужно открыть исходный файл и поправить те данные, которые программа не понимает. Распространенные проблемы:
    1. Могут дублироваться номера персон или идти не по порядку;
    2. В номерах могут быть лишние знаки;
    3. В номере персоны может быть ошибочный номер предка;
    4. Перед датами рождения/смерти могут отсутствовать стандартные знаки «\*» и «+».
    5. Даты могут быть неправильного формата (частичные обрабатываются, но дата пока обрабатывается строго формата ДД.ММ.ГГГГ), могут быть классические дополнения вида «\*1901/10» - это не обрабатывается.
    6. Если после даты есть причина смерти или место рождения/смерти/проживания/комментарий – не обрабатывается.
    7. Если в конце персональной строки есть прочие текстовые данные – также могут быть проблемы.
11. Когда проблемы исправлены (можно не закрывать журнал импорта и сверяться с ним во время правок), нужно закрыть диалог импорта, закрыть без сохранения окно базы данных и повторить шаги с 2 по 10-й. Может потребоваться несколько таких проходов.
12. Предложения по улучшению возможностей импорта – принимаются :)

# Сводка по некоторым возможным проблемам

Не импортируются файлы, если:

1. Отсутствует нумерация персон, обеспечивающая вертикальную идентификацию в древе (предок – потомок), имеются только порядковые номера.
2. В одной строке с именем персоны, также указываются не только даты рождения/смерти (допустимо), но и прочая информация – возрасты и биография (недопустимо, разумный разбор на составные части реализовать невозможно).
3. Супруги, либо родители персон не выделены в отдельные строки и описываются в общем тексте биографии персоны.
4. Дети перечислены с использованием спец.нумерации в отдельном блоке с дополнительными римскими номерами, спец.знаками и дублированием личных данных. При этом, потомки могут быть указаны в спец.блоке родителя, но отсутствовать в общем последовательном списке (дочери). Браки детей могут указываться в этих-же спец.блоках, что исключает возможность их выделения.
5. Даты рождения и смерти указываются в тексте, в разных местах, что исключает приемлемый алгоритм распознавания. Можно попробовать сделать такой разбор в будущем, в рамках создания подсистемы разбора текста «без гарантий извлечения».
6. Документ сформирован с использованием вложенной таблицы, в ячейках которой перечисляется информация (может создать трудности в случаях нелинейной структуры таблицы, но данный документ обрабатывается) [Роспись из Древа Жизни].