Система работы с табличными данными с нечеткой структурой

CONS-T-OLIDO

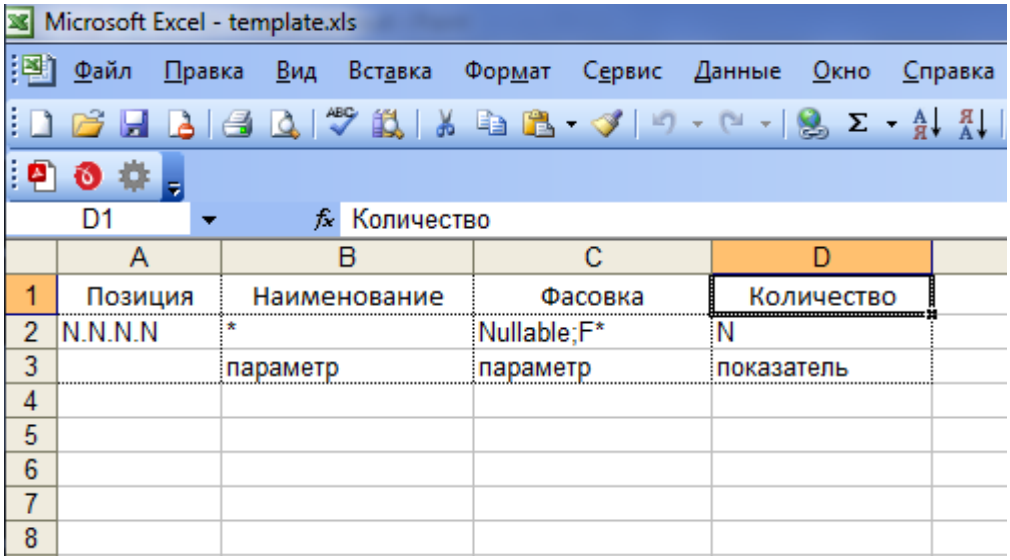
(краткое описание и начало работы)

Система работы с табличными данными с нечеткой структурой CONS-T-OLIDO (далее по тексту - Система) предназначена для экстракции данных (и, разумеется, последующей работы с ними) из источников, которые в принципе имеют общий шаблон, но в той или иной степени от него отличаются. Чаще всего это документы, которыми обмениваются сторонние организации в рамках своих деловых отношений.

Рассмотрим пример. Некая крупная организация занимается поставкой фасованных товаров розничным магазинам. Эти магазины периодически присылают заказы на поставку очередных партий товара. Наша организация на основании их заказов в свою очередь делает заказы предприятиям-изготовителям продукции. Следовательно, нам необходимо постоянно иметь итоговые цифры по каждой номенклатуре. А для отгрузки или доставки нам нужно иметь цифры по каждой номенклатуре в разрезе заказчиков.

Поскольку наши заказчики — самостоятельные организации, мы не ожидаем от них единого формата заказа. Можно лишь требовать наличия в заказе определенных реквизитов, без которых этот заказ просто невозможно выполнить.

Имея такие реквизиты, мы можем сделать приблизительный *шаблон* заказа. Для этого воспользуемся MS Excel, как одной из самых распространенных офисных программ.



	A	B	C	D
1	Позиция	Наименование	Фасовка	Количество
2	N.N.N.N	*	Nullable;F*	N
3		параметр	параметр	показатель
4				
5				
6				
7				
8				

рис.1. Пример шаблона

Определимся, что для нашего примера в заказе должны присутствовать следующие реквизиты: Позиция, Наименование, Фасовка и Количество (рис.1).

Нужно понимать, для чего мы создаем этот шаблон. Эта цель — это получить

возможность автоматической обработки документов на его основе, а не заставить сторонние организации подстраиваться под нас! Системе нужно сказать, какую информацию мы ожидаем в каждой колонке. Для этого мы указываем, во-первых, названия столбцов шаблона, а во вторых — формат их содержимого. Формат указывается *масками*.

В нашем примере мы считаем, что позиция — это номер по порядку, но с возможностью подпунктов (1, 2, 2.1, 2.2, 3, 3.1, 3.1.1 и т.д.) Для него мы принимаем маску N.N.N.N (натуральные числа с разделителями, не более 4 уровней). Значение столбца "Наименование", очевидно, может быть любым (отсюда — звездочка). Колонка "Фасовка" сочетает в себе количество (вещественное число) и единицу измерения (напр., 200 грамм, 1.5 кг). Соответственно ее маска — F*. Столбец "Количество" явно обязательный, это натуральное число (N). Также предположим, что колонка "Фасовка" необязательна для заполнения (по умолчанию — единица товара).

Из заказов наших контрагентов нам нужно получить *сводную ведомость* — количество каждой номенклатуры в разрезе заказчиков и итоги по каждой номенклатуре. Для получения такого свода в каждом реквизите шаблона указывается, параметр это или показатель. В нашем случае параметром является название номенклатуры — это колонки "Наименование" + "Фасовка" (напр., "Колбаса докторская, 300 грамм"), а показателем — столбец "Количество". Реквизиты, для которых не указано, параметр это или показатель, в сводную ведомость не попадают (но могут быть использованы для дополнительного анализа).

Как было сказано выше, ожидать идеального заполнения шаблона заказчиками весьма опрометчиво. В нашем примере сотрудники ООО "Соловушка" перепутали местами колонки "Фасовка" и "Наименование", зато добавили в заказ дополнительные листы; заодно они забыли указать номер по порядку для голландского сыра, а в третьей позиции допустили опечатку. А ИП Иванов, обладая грамотностью и аккуратностью, добавил для красоты пустую колонку между Наименованием и Фасовкой, также он сделал группировку товаров по категориям. Кроме того, в обоих заказах колонки называются не так, как в шаблоне. Но для Системы такие вольности — не проблема.

Итак, у нас есть шаблон, есть документы, более или менее ему соответствующие. Теперь нам нужен сводный отчет на их основе. Получить его из подобных заказов с помощью Системы легко: нужно просто загрузить их на сервер в в виде архива в формате zip. Архив должен содержать документы для консолидации, шаблон этих документов (имя шаблона должно быть template.xls) и т.н. *манифест* (с обязательным названием manifest.txt). В манифесте указывается, какие листы каких книг нужно обрабатывать. Пример заполнения такого манифеста:

```
--только лист Заказ
соловушка.xls:Заказ
--все листы
ип_иванов.xls:*
--все листы, кроме Доставка
ип_петров.xls:-Доставка
```

Обращаем внимание, что имена файлов и листов — регистрозависимые!

Создав архив, заходим на сайт (рис.2) и загружаем наш архив.

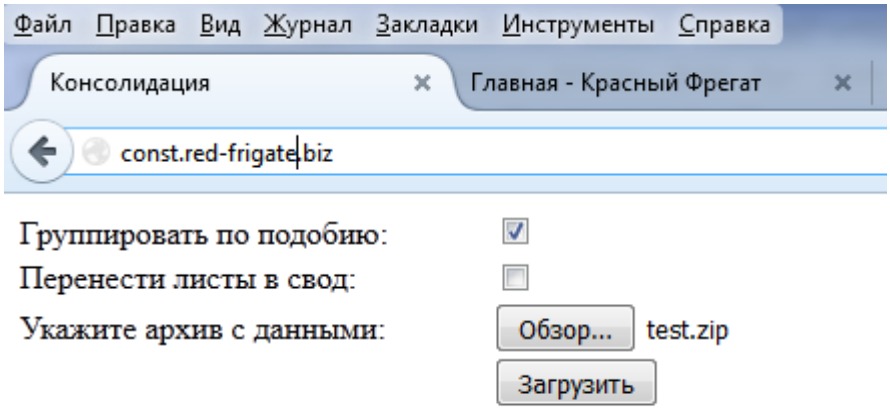


рис.2. Вид сайта консолидации

Флажок "Группировать по подобию" требует от Системы создавать группирующие записи с учетом эвристики (анализируются опечатки/ошибки и перестановки слов, здесь это "**Клбаса** сырокопченая" и "Сырокопченая колбаса"). При установленном флажке "Перенести листы в свод" в отчет будут добавлены те листы, из которых он собран.

После нажатия кнопки "Загрузить" мы получим архив, который содержит два файла. Один из них — просто сводный отчет. Другой (рис.3) — свод с группировкой; он создается, только если установлен флажок "Группировать по подобию".

	A	B	C	D	E	F	G
1	Наименование	ип_иванов.xls (ул. Дорожная)	ип_иванов.xls (ул. Садовая)	ип_петров.xls (ул. Лесная)	ип_петров.xls (ул. Луговая)	соловушка.xls (Заказ)	Всего
2	ВОДА 5Л	100	100	0	0	0	200
3	ВОДА ГАЗИРОВАННАЯ РОДНИК 1,5Л	0	0	0	0	25	25
4	ВОДА ПИТЬЕВАЯ 5Л	0	0	0	0	50	50
5	ГРУДИНКА СЫРОКОПЧЕНАЯ, В/У 300Г	25	10	0	0	0	35
6	**КЛБАСА ДОКТОРСКАЯ ДМИТРОГОРСКАЯ 0,5КГ**	**50**	**30**	**0**	**0**	**20**	**100**
7	*КЛБАСА ДОКТОРСКАЯ ДМИТРОГОРСКАЯ 0,5КГ*	*0*	*0*	*0*	*0*	*20*	*20*
8	*КОЛБАСА ДОКТОРСКАЯ ДМИТРОГОРСКАЯ 0,5КГ*	*50*	*30*	*0*	*0*	*0*	*80*
9	КОЛБАСА КРАКОВСКАЯ 0.5 КГ	30	50	0	0	0	80
10	КОЛБАСА С\К «БРАУНШВЕЙГСКАЯ» 0.5КГ	10	20	0	0	0	30

рис.3. Пример сводного отчета с группировкой

Обязательно обратите внимание на то, как Система обработала опечатку (Колбаса докторская Дмитрогорская)! Дело в том, что при анализе строк в общем случае программа не знает, какое из значений является правильным; она лишь учитывает часто встречающиеся грамматические ошибки ("о/а" в безударных

гласных, глухие/звонкие согласные и т.д.). Для повышения вероятности правильного отображения номенклатуры используются т.н. *канонизаторы*, в которых прописываются правила формирования названий для какой-то определенной предметной области. Например, в инженерных спецификациях принято писать: "Труба стальная горячекатаная в теплоизоляции ППУ Ду 50" — то есть сначала вид изделия (труба), затем его материал (стальная), после— способ изготовления (горячекатаная), потом — дополнительные характеристики (в теплоизоляции ППУ) и, наконец, для труб — диаметр. В настоящее время нами разработаны канонизаторы для большинства стройматериалов. Любые другие канонизаторы могут быть реализованы по заказу зарегистрированного пользователя системы в рамках договора сопровождения.

Сводный отчет получен! Разумеется, мы не можем знать предпочтений всех заказчиков на тему его формата и постарались реализовать стандартный свод, сообразуясь со здравым смыслом. Но, разумеется, в рамках договора сопровождения может быть доработан формат сводного отчета — для самых притязательных пользователей.

Архив с шаблоном, примерами заказов и манифестом, предназначенный для получения сводного отчета, выложен для скачивания рядом с этим документом на нашем сайте.