



**КРАСНЫЙ
ФРЕГАТ**

Primum agere

Консолидация данных с нечеткой структурой

CONS-T-OLIDO

Московская обл., г. Дубна, ул. Университетская, 9
+7(985) 885-58-13



Необходимо: получить сводные данные из источников разных форматов и с нечеткой структурой с одновременной коррекцией ошибок. Полученные данные должны быть удобны для дальнейшего анализа

➤ **Часто организации приходится обрабатывать данные в электронном виде, поступившие из внешних источников (сторонних организаций). Это могут быть сметы разных строительных организаций, сверочные ведомости от поставщиков и заказчиков...**





Проблемы обработки данных из внешних источников

➤ В силу того, что в каждой организации собственный стиль оформления документов и свой внутренний стандарт электронных средств - автоматизированная обработка таких данных стандартными инструментами (Excel, Access...) весьма и весьма затруднена.



MS Word



MS Excel

➤ Далеко не для всех товаров, услуг, работ и т.д. существует стандартное наименование, а если оно и есть – то далеко не все следуют стандартам.

Труба стальная Ду 50 ⇔ Стальные трубы 50

➤ Грамматические ошибки, опечатки, сокращения - еще больше усложняют дело

Отводы крутоизогнутые оц. Ду 50 ⇔ Отвод кр-из. оцинкованный Ф50



Пример: спецификации от двух разных подрядчиков по схожим проектам

№ п.п.	Наименование, характеристика оборудования и материалов	Тип, марка, обозначение	Единица измерения	Код оборудования	Количество	Масса, единицы оборудования кг.
1	Комплект теплосчетчика КМ-5-2	(0.03 - 24.0 м3/ч)	шт.		50	
	модуль КМППС Д 32 (0.03 - 24.0 м3/ч)		шт.		1	
	комплект монтажных частей Ду 32		шт.		1	
2	Термометр сопротивления L=80мм	КТС-Б	компл.		1	
3	Датчик давления	ИД-16	шт.		2	
4	Термометр биметаллический 0-200 °С		шт.		2	
5	Манометр технический 0-1,6 МПа	"КФV"	шт.		4	
6	Кран шаровой Ду 150	"СТС"	шт.		4	
7	Кран шаровой Ду 15 (трехходовой)	"FIV"	шт.		12	
8	Кран шаровой Ду 15	"FIV"	шт.		2	
9	Труба стальная электросварная 150	ГОСТ 10704-91	м		2	
10	Труба стальная электросварная 100	ГОСТ 10704-91	м		0,5	
11	Труба стальная электросварная 50	ГОСТ 10704-91	м		0,5	
12	Труба стальная электросварная 32	ГОСТ 10704-91	м		1	
13	Переход конический 150/100	ГОСТ 17378-83	шт.		4	
14	Переход конический 100/50	ГОСТ 17378-83	шт.		4	
15	Переход конический 50/32	ГОСТ 17378-83	шт.		4	
16	Закладная конструкция под термометр	Lпогр.= 0.3-0.7 Ду	шт.		4	
17	Закладная конструкция под КТС-Б	Lпогр.= 0.3-0.7 Ду	шт.		4	
18	Штуцер под установку манометра				4	
19	Блок питания	БП-5Б			1	
20	Адаптер периферии	АП-5Б			1	
21	Автомат электрический 6А				1	
22	Катушка имитатор Ду 32				2	
23	Провод	STR-4ST x 0,22			50	
24	Провод	ШВВП 2 x 0,5			50	
25	Провод	ПВ-3 1 x 4,0			50	
26	Металлоручая	РЗ-ЦХ10			5	
27	Гофрированная труба	ПЭ 16			100	

Изм.	Кол.уч.	Лист	№ док.	Подп.	Дата	Имя, должность	Подп. и дата
Разраб							
Проверил							

Поз.	Наименование, характеристика оборудования и материалов	Тип, марка обозначение оборудования	Единица измерения	Кол.
50	Счетчик-расходомер Ду 32	PM-5-T	шт.	1
52	комплект монтажных частей Ду 32		шт.	1
53	Датчик давления	ИД-16	шт.	1
54	Манометр технический 0-1,6 МПа		шт.	1
55	Кран шаровой ф/ф Ду 32	СТС	шт.	3
56	Кран шаровой Ду 15 (с воздухоотводчиком)	"FIV"	шт.	2
57	Кран шаровой Ду 15	"FIV"	шт.	1
58	Труба стальная электросварная оц 32	ГОСТ 10704-91	м	2
59	Отводы круглошовные о 32	ГОСТ 3262-75	шт.	2
60	Штуцер под установку манометра		шт.	1
61	Блок питания	БП-5Б	шт.	1
62	Автомат электрический 6А		шт.	1
63	Катушка имитатор Ду 32		шт.	1
64	Провод	STR-4ST x 0,12	м/п	50
65	Провод	ШВВП 2 x 0,5	м/п	50

Изм.	Кол.уч.	Лист	№ док.	Подп.	Дата	Имя, должность	Подп. и дата
Разраб							
Проверил							

Страница 1

Страница 2

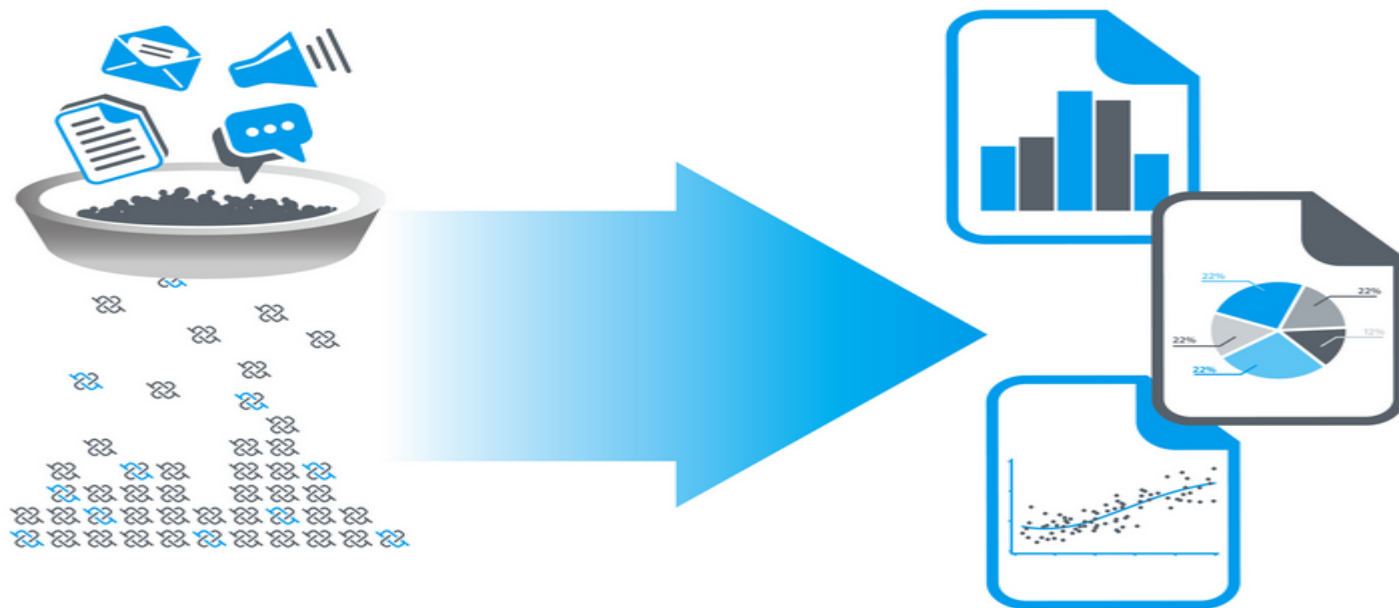


**КРАСНЫЙ
ФРЕГАТ**

Primum agere

CONS-T-OLIDO

**-система распознавания и консолидации данных
с нечеткой структурой**





Описание системы консолидации данных с нечеткой структурой **CONS-T-OLIDO**

Основа для консолидации

- **шаблон**, создаваемый пользователем.

В шаблоне указываются названия колонок и маски для них. На основании этого шаблона система распознает структуру источников данных и сами данные.

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

	A	B	C	D	E	F
1						
2	Список товаров магазина ООО «Соловушка»					
3						
4	№ п/п	Фасовка	Товар		Кол.	
5	1	200г	Сыр Российский		75	
6		200г	Сыр Голландский		20	
7	2	0,5кг	Колбаса С\к «Витязь»		10	
8	3	0,5кг	Клбаса Докторская Дмитрогорская		50	
9	4	0,5кг	Колбаса Сервелат Московский		40	
10	5	0,4кг	Хлеб Белый Нарезной		100	
11	6	0,4кг	Хлеб Черный Дарницкий		50	
12	7	300г	Шпроты Забава		20	
13	8	150г	Шпроты Забава		30	
14	9		Пакет Майка		300	
15	10		Пакет		300	
16						
17	11	5л	Вода		100	
18	12	1,5л	Вода Газированная		50	
19						
20	Ответственный: Иванов П.С.					
21						



	A	B	C	D	E	F
1						
2	Поз.	Наименование товара		Фасовка	Кол-во	
3	1	Сыр Российский		200г	40	
4	1,1	Сыр Голландский		200г	25	
5	2	Колбаса С\к «Витязь»		0,5кг	20	
6	3	Колбаса Докторская Дмитрогорская		0,5кг	20	
7	4	Колбаса Сервелат Московский		0,5кг	20	
8	5	Хлеб Белый Нарезной		0,4кг	45	
9	6	Хлеб Черный Дарницкий		0,4кг	15	
10	7	Шпроты Забава		300г	50	
11	9	Пакет Майка			300	
12	11	Вода		5л	15	
13	12	Вода Газированная		1,5л	25	
14						
15						
16						

	A	B	C	D	E
1	Поз.	Товар	Фасовка	Кол-во	
2	1	Российский Сыр	200г	100	
3		Голландский Сыр	200г	100	
4	2	С\к Колбаса «Витязь»	0,5кг	100	
5	3	Докторская Колбаса Дмитрогорская	0,5кг	50	
6	4	Колбаса Сервелад московск	0,5кг	50	
7	5	Нарезной Хлеб Белый	0,4кг	50	
8	6	Дарницкий Хлеб Черный	0,4кг	40	
9	7	Забава Шпроты	300г	30	
10	8	Забава Шпроты	150г	50	
11	9	Майка Пакет		200	
12	10	Пакет		200	
13	12	Газированная Вода	1,5л	130	
14					
15					
16					



Описание системы консолидации данных с нечеткой структурой **CONS-T-OLIDO**

Результат работы системы – **сводный «шахматный» отчет**, в котором для схожих позиций создаются группирующие записи.

	A	B	C	D	E
	Наименование	ИП	ООО	ООО	Всего
1		Козлов	Агат	Соловушка	
2	ВОДА 5Л	0,	15,	100,	115,
3	ВОДА ГАЗИРОВАННАЯ 1,5Л	130,	25,	50,	205,
4	<i>ВОДА ГАЗИРОВАННАЯ 1,5Л</i>	0,	25,	50,	75,
5	<i>ГАЗИРОВАННАЯ ВОДА 1,5Л</i>	130,	0,	0,	130,
6	КОЛБАСА ДОКТОРСКАЯ ДМИТРОГОРСКАЯ 0,5КГ	50,	20,	50,	120,
7	<i>КЛБАСА ДОКТОРСКАЯ ДМИТРОГОРСКАЯ 0,5КГ</i>	0,	0,	50,	50,
8	<i>ДОКТОРСКАЯ КОЛБАСА ДМИТРОГОРСКАЯ 0,5КГ</i>	50,	0,	0,	50,
9	<i>КОЛБАСА ДОКТОРСКАЯ ДМИТРОГОРСКАЯ 0,5КГ</i>	0,	20,	0,	20,
10	КОЛБАСА С\К «ВИТЯЗЬ» 0,5КГ	100,	20,	10,	130,
11	<i>КОЛБАСА С\К «ВИТЯЗЬ» 0,5КГ</i>	0,	20,	10,	30,
12	<i>С\К КОЛБАСА «ВИТЯЗЬ» 0,5КГ</i>	100,	0,	0,	100,
13	КОЛБАСА СЕРВЕЛАТ МОСКОВСКИЙ 0,5КГ	50,	20,	40,	110,
14	<i>КОЛБАСА СЕРВЕЛАТ МОСКОВСКИЙ 0,5КГ</i>	0,	20,	40,	60,
15	<i>КОЛБАСА СРВЕЛАД МОСКОВСК 0,5КГ</i>	50,	0,	0,	50,
16	ПАКЕТ	200,	0,	300,	500,
17	ПАКЕТ МАЙКА	200,	300,	300,	800,
18	<i>ПАКЕТ МАЙКА</i>	0,	300,	300,	600,
19	<i>МАЙКА ПАКЕТ</i>	200,	0,	0,	200,
20	СЫР ГОЛЛАНДСКИЙ 200Г	100,	25,	20,	145,
21	<i>СЫР ГОЛЛАНДСКИЙ 200Г</i>	0,	25,	20,	45,
22	<i>ГОЛЛАНДСКИЙ СЫР 200Г</i>	100,	0,	0,	100,
23	СЫР РОССИЙСКИЙ 200Г	100,	40,	75,	215,
24	<i>СЫР РОССИЙСКИЙ 200Г</i>	0,	40,	75,	115,
25	<i>РОССИЙСКИЙ СЫР 200Г</i>	100,	0,	0,	100,
26	ХЛЕБ БЕЛЫЙ НАРЕЗНОЙ 0,4КГ	50,	45,	100,	195,
27	<i>ХЛЕБ БЕЛЫЙ НАРЕЗНОЙ 0,4КГ</i>	0,	45,	100,	145,
28	<i>НАРЕЗНОЙ ХЛЕБ БЕЛЫЙ 0,4КГ</i>	50,	0,	0,	50,
29	ХЛЕБ ЧЕРНЫЙ ДАРНИЦКИЙ 0,4КГ	40,	15,	50,	105,
30	<i>ХЛЕБ ЧЕРНЫЙ ДАРНИЦКИЙ 0,4КГ</i>	0,	15,	50,	65,
31	<i>ДАРНИЦКИЙ ХЛЕБ ЧЕРНЫЙ 0,4КГ</i>	40,	0,	0,	40,
32	ШПРОТЫ ЗАБАВА 150Г	50,	0,	30,	80,
33	<i>ШПРОТЫ ЗАБАВА 150Г</i>	0,	0,	30,	30,
34	<i>ЗАБАВА ШПРОТЫ 150Г</i>	50,	0,	0,	50,
35	ШПРОТЫ ЗАБАВА 300Г	30,	50,	20,	100,
36	<i>ШПРОТЫ ЗАБАВА 300Г</i>	0,	50,	20,	70,
37	<i>ЗАБАВА ШПРОТЫ 300Г</i>	30,	0,	0,	30,
38					



Описание системы консолидации данных с нечеткой структурой **CONS-T-OLIDO**

Группировка – с учетом грамматических ошибок, опечаток и сокращений – осуществляется на основе правил русского языка. Однако для конкретной предметной области могут быть сделаны специальные функции канонизации.

	A	B	C	D	E	F
1	Наименование	Горбачево Мастерская ПЧ ХВС	Казакова, 8 ЦТП ЦО (Мосэнерго)	Коломенское МОЭК ТЭ	Всего	
2	КАТУШКА ИМИТАТОР ДУ 32	1,	2,	4,	7,	
3	КОМПЛЕКТ МОНТАЖНЫХ ЧАСТЕЙ ДУ 15	0	0	1	1	
4	КОМПЛЕКТ МОНТАЖНЫХ ЧАСТЕЙ ДУ 32	1,	1,	1,	3,	
5	<i>КОМПЛЕКТ МОНТАЖНЫХ ЧАСТЕЙ 32</i>	<i>1,</i>	<i>0,</i>	<i>0,</i>	<i>1,</i>	
6	<i>КОМПЛЕКТ МОНТАЖНЫХ ЧАСТЕЙ Ф 32</i>	<i>0,</i>	<i>1,</i>	<i>0,</i>	<i>1,</i>	
7	<i>КОМПЛЕКТ МОНТАЖНЫХ ЧАСТЕЙ ДУ 32</i>	<i>0,</i>	<i>0,</i>	<i>1,</i>	<i>1,</i>	
8	КОМПЛЕКТ ТЕПЛОСЧЕТЧИКА ГВС КМ-5-4	0,	0,	1,	1,	
9	КОМПЛЕКТ ТЕПЛОСЧЕТЧИКА КМ-5-2 (0.03 - 24 МЗ/Ч)	0,	50,	0,	50,	
10	КОМПЛЕКТ ТЕПЛОСЧЕТЧИКА ОТОПЛЕНИЯ КМ-5-2	0,	0,	1,	1,	
11	КРАН ШАРОВОЙ ДУ 15 (С ВОЗДУХООТВОДЧИКОМ) FIV	2,	0,	13,	15,	

Кроме того, для заданной предметной области также могут быть разработаны специальные шаблоны.



Описание системы консолидации данных с нечеткой структурой **CONS-T-OLIDO**

Система имеет различные варианты использования и комплекты поставки:

- **Веб-приложение:** на сайт загружается zip-архив, состоящий из шаблона и таблиц с данными. В ответ пользователь получает тоже архив: сводный отчет и лог с ошибками обработки.
- **Настольное кросс-платформенное приложение, стандартная редакция** – для тех, кто не доверяет свои данные кому бы то ни было. Система запускается из командной строки, указывается файл архива с шаблоном и данными. На выходе – опять же сводный отчет и лог ошибок обработки.
- **Настольное кросс-платформенное приложение, расширенная редакция** – рекомендуется тем, кому необходимы дополнительные анализ и обработка данных. Как исходные, так и консолидированные данные загружаются в базу данных, после чего с ними можно работать различными инструментами (построитель отчетов «Москит», средства для работы с многомерными данными **Red OLAP Navigator**, прикладные программы ODBC/JDBC...) В этом случае в комплект поставки входит СУБД «**РЕД** База Данных», открытая или enterprise редакция.



**КРАСНЫЙ
ФРЕГАТ**

Primum agere

Спасибо за внимание!