ETL процессы

Александр Дольник

alexanderdolnik@acm.org

-Ты кто по профессии?
-Я? Магистр черной и белой бухгалтерии!
Из анекдотов с бородой

План 5-минутки

- Метаданные предметной области
- Общий список задач
- Примеры по теме
- Демо

Метаданные предметной области

- ***Метаданные -*** данные описывающие данные: каталоги, справочники, реестры, базы метаданных, содержащие сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.
- ***Онтология*** (ontology) - это попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной схемы
- ***ETL процесс***

Что такое ETL?

- **ETL** (от англ. *Extract, Transform, Load* — дословно «извлечение, преобразование, загрузка») — один из основных процессов в управлении хранилищами данных, который включает в себя:
  - извлечение данных из внешних источников;
  - их трансформация и очистка, чтобы они соответствовали нуждам бизнес-модели;
  - и загрузка их в хранилище данных.

Общий список задач

- **Task 1** Нечёткое отображение данных с опорой на онтологию – идентификация понятий предметной области
  - ***Пример***: контрольные работы

- **Task 2** Эволюция связей между/внутри электронных таблиц и идентификация изменений в табличных данных (вставка, модификация, удаление)

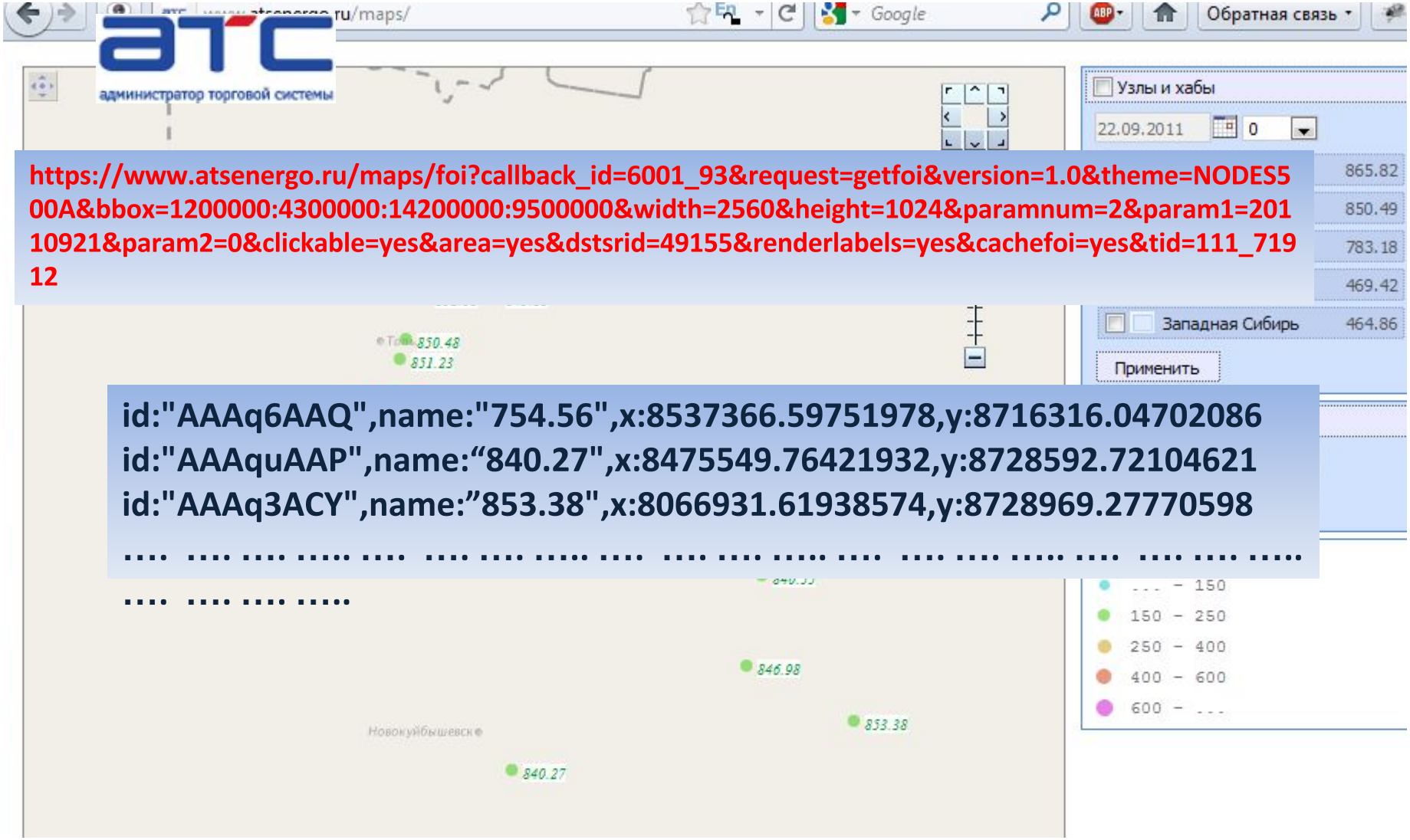
- **Task 3** Связывание различных данных одной предметной области с использованием онтологий и журнала изменений
  - ***Пример***: АТС (администратор торговой системы)

Task 1. Контрольные работы

Task 1. Контрольные работы (прод.)

Здесь уже необходима дополнительная информация … ??? ОнтологиЯ ???

Task 3. Пример. Администратор торговой системы



Task 3. Пример. Администратор торговой системы

Узел	Код ЗСП	Наименование ЗСП
100421	FZURPZ11	Пермско-Закамский энергоузел Пермской энергосистемы
100422	FZURPZ11	Пермск
100423	FZURPZ11	Пермск
…	…	

Новые узлы		Удаленные узлы	
Узел	**Код ЗСП**	**Узел**	**Код ЗСП**
100286	FZUROE07	510056	FZZMSK26
100477	FZUROE07	510342	FZZMSK26
100479	FZUROE07	510343	FZZMSK26
100850	FZUROE07	510536	FZZMSK26
100852	FZUROE07	510537	FZZMSK26
	…	…	…

Изменения – могут помочь в установлении географического отображения?

Демонстрация силы

- http://www.youtube.com/watch?v=wZMh7uF0_QE
- **Data Normalisation with Kettle / Pentaho Data Integration**

Бонусы данной области

- Это не абстрактная наука, но прикладная область
- Находится на грани SE и CS
  - Есть вопросы касающиеся как дизайна, архитектуры,
  - Так и связанные с разработкой новых алгоритмов, которые способны корректно (гарантировано) обработать большой объём данных за приемлемое время
- Может затрагивать очень многие области информатики
  - Такие как: теорию БД, теорию графов, теорию информации, мат. статистику, дискретную математику и другие
- Это действительно интересно!