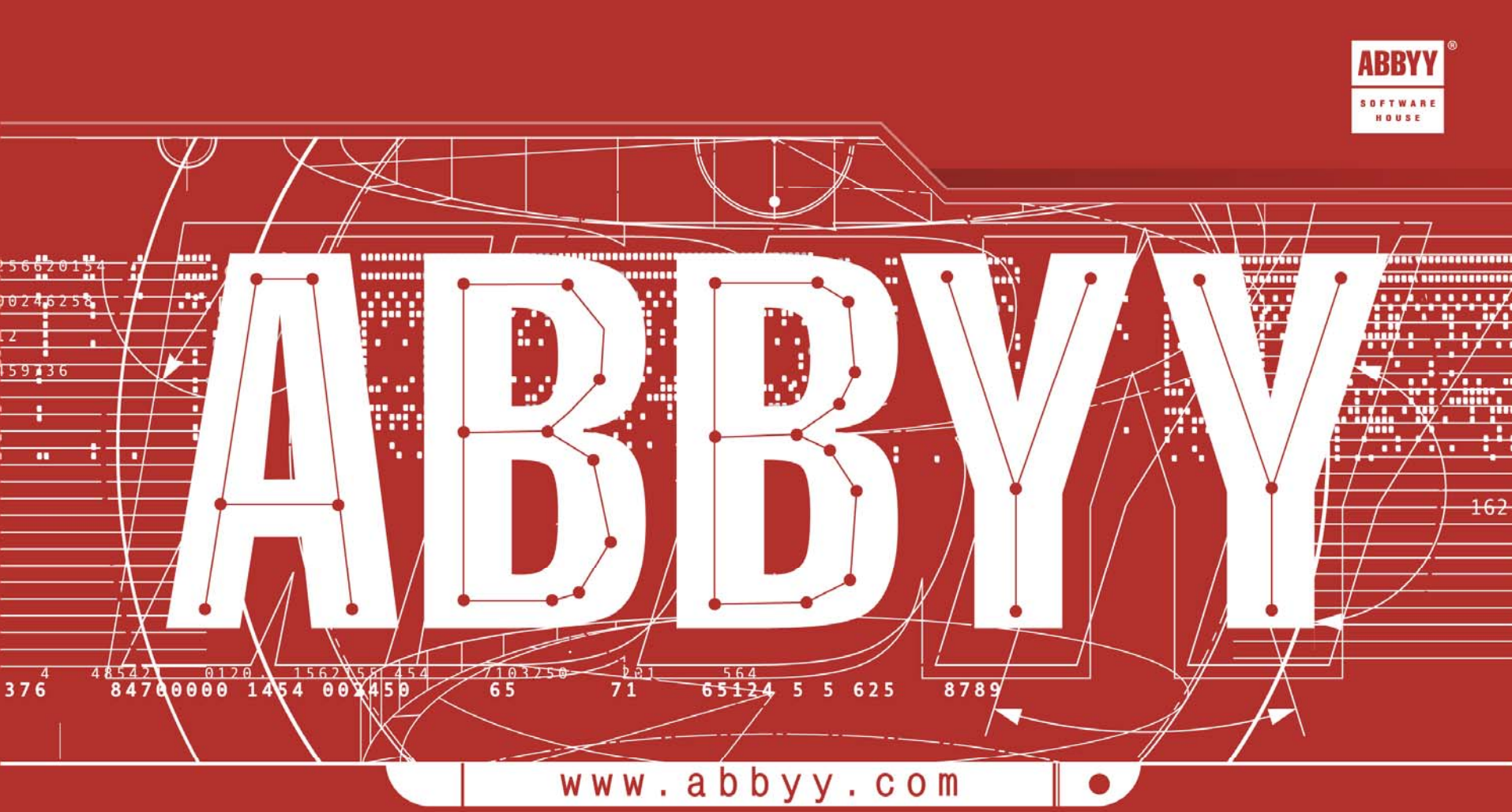


Проблемы анализа данных – подходы к решению задач нормализации и идентификации.

Диар Туганбаев, Руководитель группы извлечения знаний из текста
Андрей Лубенец, Менеджер проектов по интеграции технологий

Проблемы анализа данных

ИНН

7734567540 ? 7734567540

АДРЕС

М.О. Долгопрудный ул Первомайская 5 11 ? г. Долгопрудный Мытищинского района Московской области Первомайская 5 11

ЧТО ДЕЛАТЬ?

Прикладные задачи

- Внедрение ERP / DMS / CMS / CRM – загрузка данных
- Слияние баз информационных систем (например, при объединении компаний)
- Отслеживание повторений в справочниках

Основная задача

- **Предоставить возможность использования текстовой информации для анализа**
  - Адреса
  - Наименования организаций, банков
  - Описания операций
  - ФИО
  - Место рождения
  - …

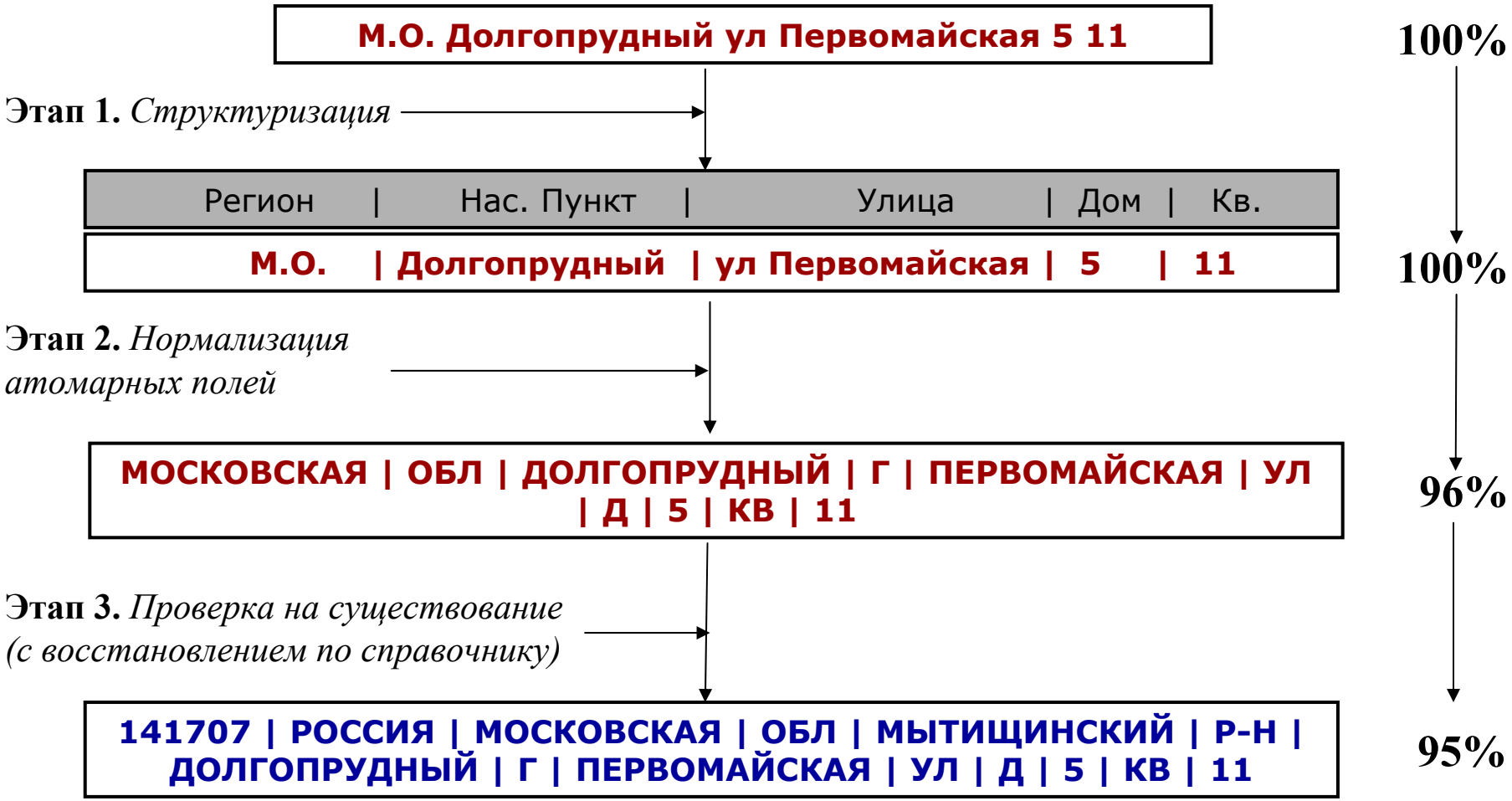
Что такое нормальный вид?

- Нормальный вид типа данных:
  - Основной критерий – возможность простым сравнением любых нормализованных объектов одного типа сказать – равны эти объекты или нет.
  - Нормальный вид типа данных определяется аналитиком и разработчиком в контексте прикладной задачи
- Нормальный вид атомарного типа данных:
  - Определяется по словарю или регулярному выражению

Подзадачи нормализации

- Структуризация – разбиение сводного поля на атомарные поля
- Приведение атомарных полей к нормальному виду
- Проверка объекта на существование (опциональный этап)
- Восстановление /обогащение данных (опциональный этап)

Нормализация адреса



Структуризация

Москва	**Ямского**	**поля**	**5**	**10**
Регион	*Город*	*Улица*	*Дом*	*Квартира*
Город	*Улица*		*Дом*	*Квартира*
Город		*Улица*	*Дом*	*Квартира*
Город	*Улица*			*Дом*

Сравнительная важность подзадач для разных типов данных

Тип данных	Структуризация	Нормализация атомарных полей	Проверка на существование
БИК или SWIFT код	–	+	+
Адрес	+	+	+
ФИО	+	+	–

```
<?xml version="1.0" ?>
- <AddressDescription MinQuality="0" SecondStageQualityThreshold="80">
  - <ImportedDescriptions>
      <StandardComponents FileName="standardComponents.xml" ImportRegularExpressions="yes" ImportSimpleComponents="yes"
        ImportResultComponents="yes" ImportTranslations="yes" ImportDictionaries="yes" />
    </ImportedDescriptions>
  - <Dictionaries>
      <RegionNamesDictionary FileName="AllRegionNames.txt" />
      <DistrictNamesDictionary FileName="AllDistrictNames.txt" />
      <DistrictQualifiersDictionary FileName="districtQualifiers.txt" />
      <CityNamesDictionary FileName="AllCityNames.txt" />
      <SettlementNamesDictionary FileName="AllSettlementNames.txt" />
      <StreetNamesDictionary FileName="AllStreetNames.txt" />
      <StreetQualifiersDictionary FileName="streetQualifiers.txt" />
      <HouseQualifiersDictionary FileName="houseQualifiers.txt" />
      <FlatNamesDictionary FileName="flatNames.txt" />
      <OptionalWordsDictionary FileName="optionalWords.txt" />
      <CityNamePartsDictionary FileName="cityNameParts.txt" />
      <StreetNamesPartDictionary FileName="streetNameParts.txt" />
      <MinorNamePartsDictionary FileName="minorNameParts.txt" />
      <CountryNameDictionary FileName="countryNames.txt" />
      <CityPhoneCodesDictionary FileName="cityPhoneCodes.txt" />
      <CountryPhoneCodesDictionary FileName="countryPhoneCodes.txt" />
      <CityDistrictsDictionary FileName="cityDistricts.txt" />
    </Dictionaries>
  - <RegularExpressions>
      <PhoneQualifier>"ТЕЛ"</PhoneQualifier>
      <UnknownWord>[А-ЯЁ]{2-}</UnknownWord>
      <UnknownWords>(<regexp:UnknownWord>[ \-]){-}<regexp:UnknownWord></UnknownWords>
      <Int>[1-9][0-9]{0-3}</Int>
      <Letter>[A-Z]|[А-ЯЁ]</Letter>
      <IntSuffix><regexp:Int>(" "|"-"|)([А-Я]|)</IntSuffix>
      <ComplexNumber>[1-9][0-9]{0-3}(|<regexp:Letter>)</ComplexNumber>
      <CountryZip>(<data:Rubbish>{0-1}) ((|<result:CountryName>)<result:ZipCode>)| ((|<result:ZipCode>)
        <result:CountryName>) (<data:Rubbish>{0-1})</CountryZip>
      <Number0to999>[0-9]{1-3}</Number0to999>
      <LetterOrThree>[А-ЯЁ]{1-3}</LetterOrThree>
      <Region>(<data:Rubbish>{0-1}) ((((|<result:RegionQualifier>) (|<result:OptionalRegion>)<result:RegionName>
        (|<result:OptionalRegion>))| ((|((|<result:OptionalRegion>)<result:RegionName>(|<result:OptionalRegion>)))
        <result:RegionQualifier>)) (<data:Rubbish>{0-1})</Region>
```

Проблемы анализа

ИНН

7734567540	=	7734567540

АДРЕС

141707 | РОССИЯ | МОСКОВСКАЯ | ОБЛ | МЫТИЩИНСКИЙ | Р-Н | ДОЛГОПРУДНЫЙ | Г | ПЕРВОМАЙСКАЯ | УЛ | Д | 5 | КВ | 11	≠ НОРМАЛИЗАЦИЯ	141707 | РОССИЯ | МОСКОВСКАЯ | ОБЛ | МЫТИЩИНСКИЙ | Р-Н | ДОЛГОПРУДНЫЙ | Г | ПЕРВОМАЙСКАЯ | УЛ | Д | 1 5 | КВ | 11

Достижения

- Структуризация полей свободного формата
  - "Москва" – город, "Ленинский проспект" – улица
- Универсальный подход
  - Описание типа данных отделено от энжина нормализации, т.е. возможна настройка обработки типов данных без программирования
- Нечеткое сравнение строк (с учетом опечаток)
- Транслитерация и сравнение по созвучию

Вопросы

Диар Туганбаев,
Руководитель группы извлечения знаний из текста
Diar_T@abbyy.com

Андрей Лубенец,
Менеджер проектов по интеграции технологий
Andrey_L@abbyy.com