

АВВУУ

The logo consists of the Cyrillic letters 'А', 'В', 'В', 'У', 'У' in a bold, white, sans-serif font. Each letter is overlaid with a network diagram of red dots connected by thin red lines, suggesting a data structure or neural network. The background is a dark red with white technical drawings, including lines, circles, and arrows, and some faint numerical data.

[www.abbyy.com](http://www.abbyy.com)

**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**

**ИНН**

**7734567540**



**7734567540**

**АДРЕС**

**М.О. Долгопрудный ул  
Первомайская 5 11**

**ЧТО ДЕЛАТЬ?**

**г. Долгопрудный  
Мытищинского района  
Московской области  
Первомайская 5 11**

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

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

- **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

- 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

**М.О. Долгопрудный ул Первомайская 5 11**

100%

Этап 1. Структуризация

| Регион      | Нас. Пункт          | Улица                  | Дом      | Кв.       |
|-------------|---------------------|------------------------|----------|-----------|
| <b>М.О.</b> | <b>Долгопрудный</b> | <b>ул Первомайская</b> | <b>5</b> | <b>11</b> |

100%

Этап 2. Нормализация  
атомарных полей

**МОСКОВСКАЯ | ОБЛ | ДОЛГОПРУДНЫЙ | Г | ПЕРВОМАЙСКАЯ | УЛ  
| Д | 5 | КВ | 11**

96%

Этап 3. Проверка на существование  
(с восстановлением по справочнику)

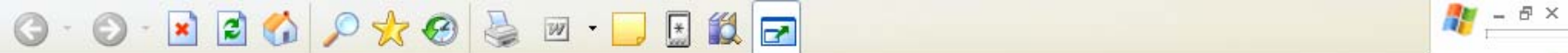
**141707 | РОССИЯ | МОСКОВСКАЯ | ОБЛ | МЫТИЩИНСКИЙ | Р-Н |  
ДОЛГОПРУДНЫЙ | Г | ПЕРВОМАЙСКАЯ | УЛ | Д | 5 | КВ | 11**

95%

|               |                |              |            |                 |
|---------------|----------------|--------------|------------|-----------------|
| <b>Москва</b> | <b>Ямского</b> | <b>поля</b>  | <b>5</b>   | <b>10</b>       |
| <i>Регион</i> | <i>Город</i>   | <i>Улица</i> | <i>Дом</i> | <i>Квартира</i> |
| <i>Город</i>  | <i>Улица</i>   |              | <i>Дом</i> | <i>Квартира</i> |
| <i>Город</i>  |                | <i>Улица</i> | <i>Дом</i> | <i>Квартира</i> |
| <i>Город</i>  | <i>Улица</i>   |              |            | <i>Дом</i>      |

# 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

НОРМАТИВНАЯ



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

- 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

Диар Туганбаев,  
Руководитель группы извлечения знаний из текста  
[Diar\\_T@abbyy.com](mailto:Diar_T@abbyy.com)

Андрей Лубенец,  
Менеджер проектов по интеграции технологий  
[Andrey\\_L@abbyy.com](mailto:Andrey_L@abbyy.com)