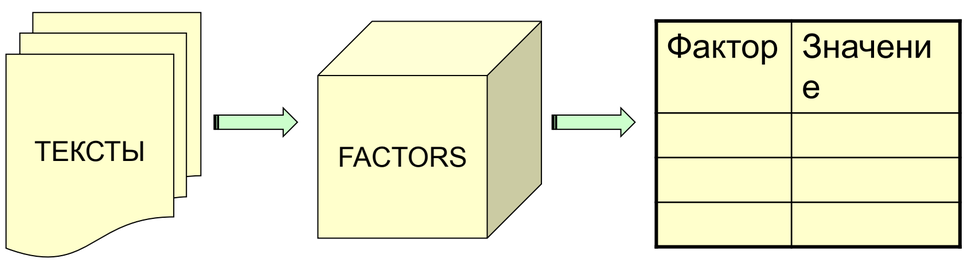
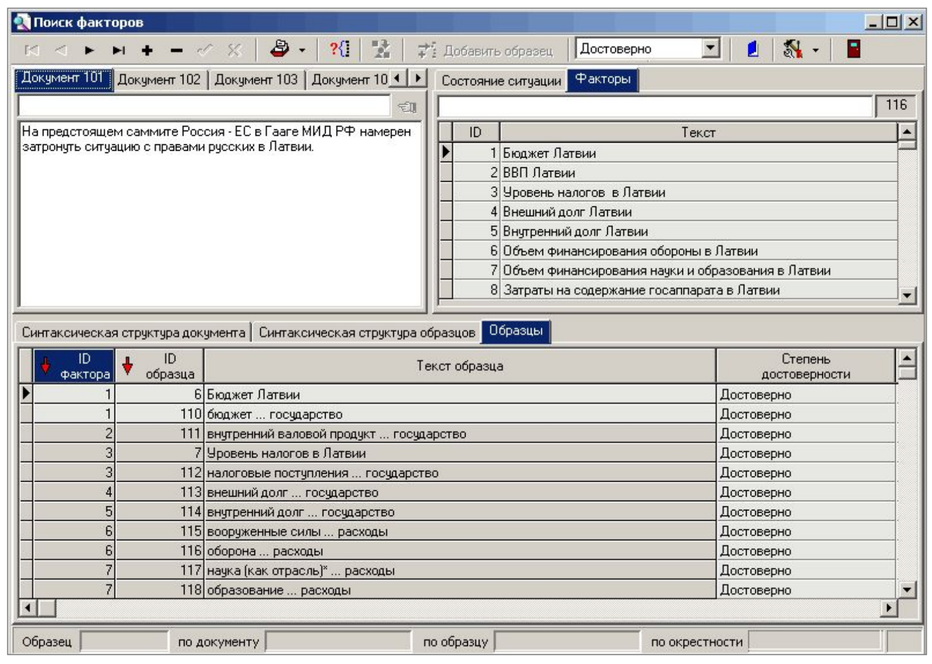
# СИСТЕМА ИЗВЛЕЧЕНИЯ ФАКТОГРАФИЧЕСКОЙ ИНФОРМАЦИИ ИЗ ТЕКСТОВ ОБЩЕСТВЕННО-ПОЛИТИЧЕСКОЙ ТЕМАТИКИ

**Пивоварова Л. М. (СПбГУ)**
Научный руководитель: Рубашкин В. Ш.

Система Factors:

- интеллектуальная среда для поддержки работы эксперта-аналитика с текстами.




**Задача**: извлечение из текстов СМИ информации общественно-политической тематики.

Факторы - различные характеристики общественно-политической ситуации (*число пенсионеров; средний уровень заработной платы; социальная напряженность; военные угрозы*).

Значения факторов: количественные (*объем экспорта*) и оценочные (*уровень плюрализма в СМИ*)

**Функциональность:**

1. Последовательное наращивание распознаваемых аспектов содержания в процессе работы эксперта-аналитика с системой.
2. Легкость и простота редактирования и пополнения; визуальное представление информации.
3. Функциональная расширяемость и переносимость на другие проблемные и предметные области.

**Методология Information Extraction**:
поиск на основе текстовых образцов.

**Образцы**

Фактор + значение

В основном для оценочных факторов

*социальная напряженность →*
*стихийный митинг*

Только фактор

Для количественных образцов:

*уровень инфляции →*
*инфляция составила 4%*

**Образцы**

Текстовые – выделение в тексте релевантных фрагментов (при анализе может проверяться совпадение синтаксических связей)

Концептуальные – сборка образца из концептов **онтологии** (при анализе осуществляется поиск с учетом отношения «общее-частное»)

Смешанные

## Поиск образцов в тексте

Собственный признак фактора – концепт, отвечающий на вопрос «количество (величина) чего?»

*Уровень зарплаты → заработная плата*
*Транспортные издержки → траты*
*Число пенсионеров → пенсионеры*

Онтология: собственный признак ↔ единица измерения

*заработная плата ↔ денежная единица*
*пенсионеры ↔ без единиц*

**Общий алгоритм поиска**

1) Поиск образца
2) Определение собственного признака и единиц измерения
3) Поиск числа с единицей измерения
4) Проверка соответствия единиц измерения
5) Если число не найдено – поиск слов *большой, маленький, растет, падает* и их синонимов
6) Определение достоверности

Параметры поиска предполагают отладку и настройку