

Программа для извлечения метаописаний
из неструктурированных текстов
METADATA EXTRACTOR

Руководство
пользователя

Версия от 2018 г.

Оглавление

Введение	3
Принцип работы.....	3
Меню.....	4
Тайтл бар	5
Вкладки.....	6
1 Control (Рис. 5)	6
2 Info (Рис. 8)	8
3 Contact (Рис. 9)	9
4 Person (Рис. 10).....	10
5 Keyword (Рис. 11).....	11
6 Location (Рис. 12).....	11
7 Reference (Рис. 13).....	12
Статус бар (Рис. 14).....	13
Помощь	13
Настройки.....	14
О программе.....	15
Работа через командную строку	16
Другие настройки.....	16
Краткая инструкция по использованию программы	17

Введение

Программа METADATA EXTRACTOR предназначена для извлечения метаописаний из неструктурированных текстов формата pdf. Она помогает частично автоматизировать процесс подготовки метаописаний для системы GeoNetwork.

METADATA EXTRACTOR извлекает данные следующих видов:

1. Заголовок документа
2. Год публикации
3. Имена
4. Локации
5. Названия организация
6. Разное
7. Библиографические ссылки
8. Ключевые слова и словосочетания

METADATA EXTRACTOR работает под управлением операционной системы Windows и Linux.

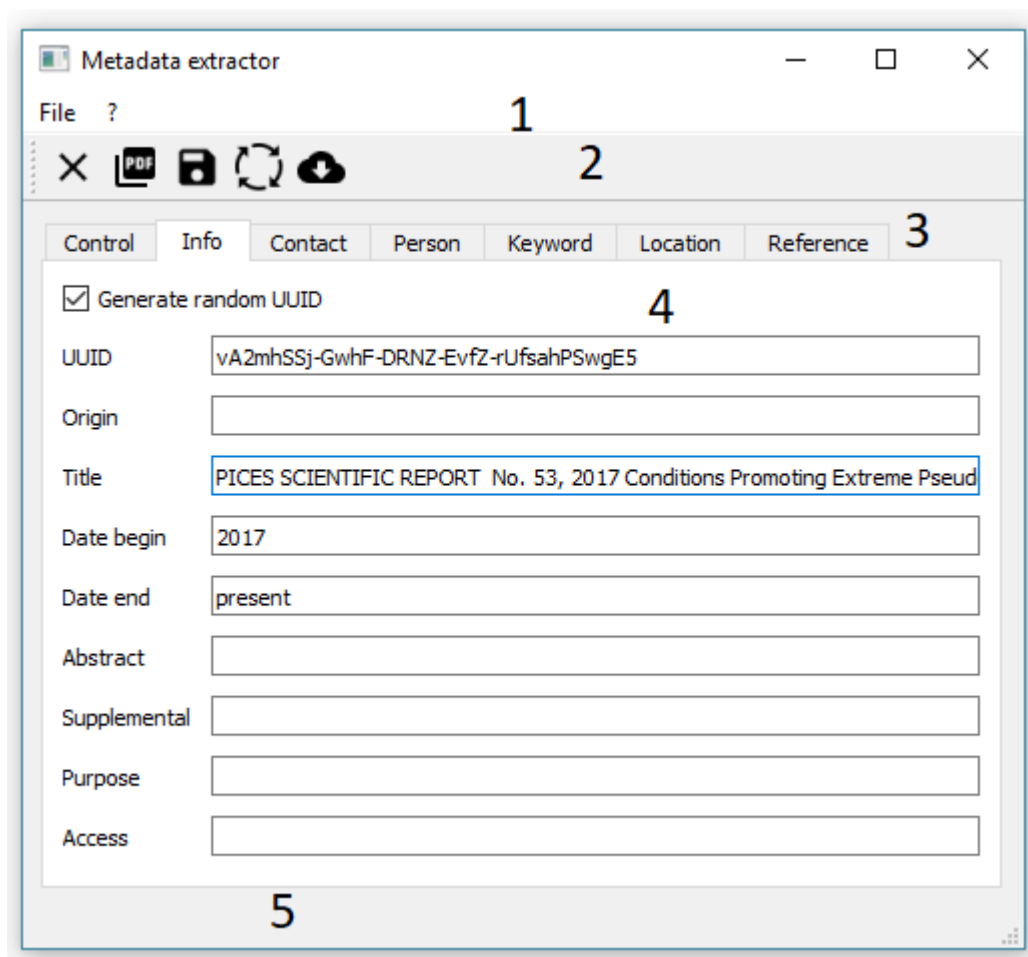
Note: тестировалось на ОС Windows 10 и GNU/Linux Manjaro 17.1.10

Принцип работы

На вход программе подается pdf файл, далее после его обработки автоматически заполняются поля форм. Извлечённые данные можно сохранить в формате txt, iso19115v2 или fgdc. Также данные можно отправить на сервер GeoNetwork.

Программа запускается с помощью *mainGUI.py* (графический режим) *main.py* (интерфейс командной строки).

Окно программы показано на Рис. 1.



“Рис. 1 Окно программы”

На Рис. 1 цифрами отмечены основные элементы программы.

- 1 – Меню – предоставляет доступ к основным функциям программы.
- 2 – Тайтл бар – быстрый доступ к отдельным функциям программы.
- 3 – Вкладки с отдельными классами данных.
- 4 – Форма активной вкладки
- 5 – Статус бар – выводит сообщения от программы.

Меню

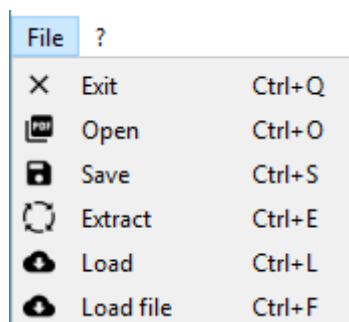
В меню два пункта: “File” и “?” (Рис. 2).

File ?

“Рис. 2 Меню”

File - Управление приложением (Рис. 3)

? – Справка и настройки (Рис. 4)

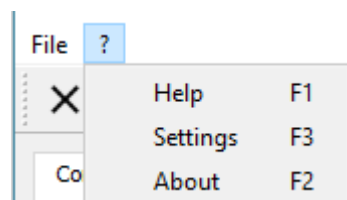


“Рис. 3 «File»”

Команды меню «File» описаны в Табл. 1.

Название	Комбинация клавиш	Действие
Exit	Ctrl+Q	Выход из приложения
Open	Ctrl+O	Открыть PDF файл
Save	Ctrl+S	Сохранить данные в формате txt, iso19115v2, fgdc
Extract	Ctrl+E	Запустить процесс извлечения метаданных из выбранного pdf файла
Load	Ctrl+L	Загрузить метаданные на сервер GeoNetwork в формате iso19115v2 или fgdc
Load file	Ctrl+F	Загрузить метаданные из существующих файлов в формате iso19115v2 или fgdc

“Табл. 1 Меню File”



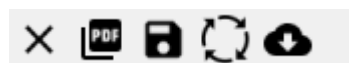
“Рис. 4 «?»”

Команды меню «?» описаны в Табл. 2.

Название	Комбинация клавиш	Действие
Help	F1	Открывает «Документацию»
Settings	F3	Открывает «Настройки»
About	F2	Открывает «О программе»


“Табл. 2 Меню ?”

Тайтл бар



“Рис. 5 «Тайтл бар»”

Команды тайтл бара описаны в Табл. 3.

Название	Комбинация клавиш	Действие
	Ctrl+Q	Выход из приложения
	Ctrl+O	Открыть PDF файл
	Ctrl+S	Сохранить данные в формате txt, iso19115v2, fgdc
	Ctrl+E	Запустить процесс извлечения метаданных из выбранного pdf файла
	Ctrl+L	Загрузить метаданные на сервер GeoNetwork в формате iso19115v2 или fgdc

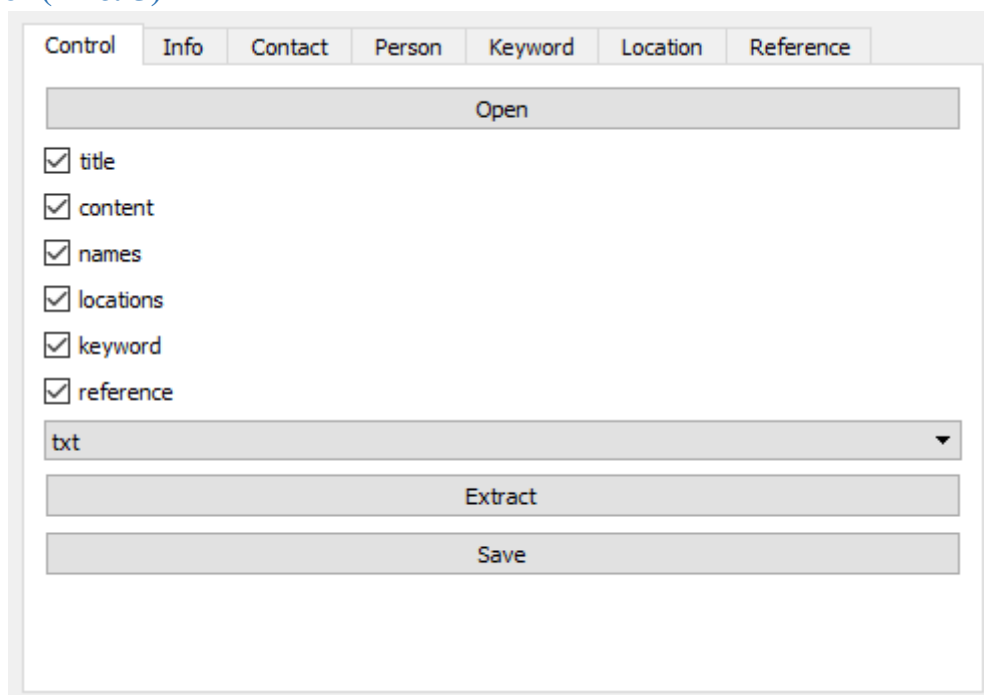
“Табл. 3 Тайтл бар”

Вкладки

Элементы приложения разбиты на вкладки:

1. Control – конфигурирует работу приложения
2. Info – основные данные
3. Contact – лицо ответственное за данные
4. Person - имена
5. Keyword – ключевые слова и словосочетания
6. Location – локации с географическими координатами
7. Reference – библиографические ссылки

1 Control (Рис. 5)



The screenshot shows the 'Control' tab of the application. At the top, there is a row of tabs: Control, Info, Contact, Person, Keyword, Location, and Reference. The 'Control' tab is active. Below the tabs, there is a large 'Open' button. Underneath the button is a list of checkboxes, each followed by a label: ☒ title, ☒ content, ☒ names, ☒ locations, ☒ keyword, and ☒ reference. Below the checkboxes is a dropdown menu with 'txt' selected. At the bottom of the tab, there are two buttons: 'Extract' and 'Save'.

“Рис. 5 «Вкладка «Control»”

Кнопка «*Open*» - вызывает диалоговое окно Рис. 6, где пользователь выбирает pdf файл и нажимает на кнопку “Open”.

Чекбокс «*title*» – если активен, то будет извлекаться заголовок.

Чекбокс «*content*» – если активен, то будет извлекаться оглавление (только для txt).

Чекбокс «*names*» – если активен, то будут извлекаться имена.

Чекбокс «*location*» – если активен, то будут извлекаться локации.

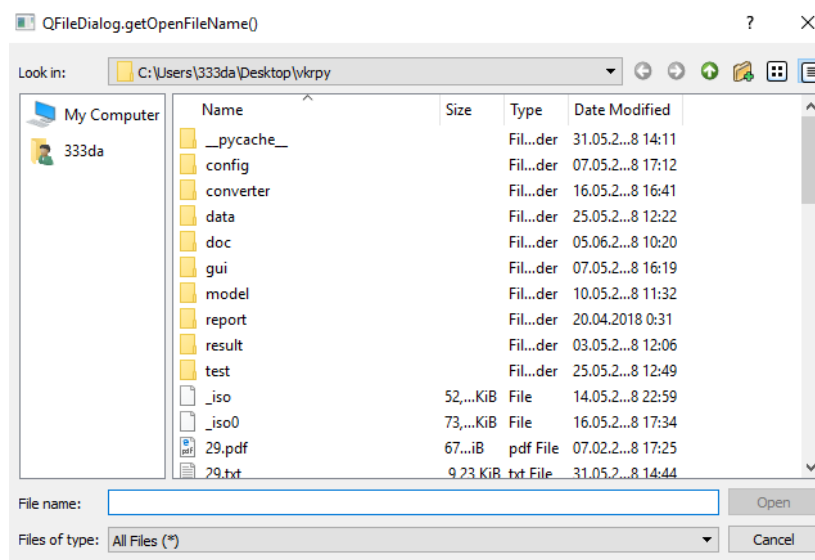
Чекбокс «*keyword*» – если активен, то будут извлекаться ключевые слова.

Чекбокс «*reference*» – если активен, то будут извлекаться библиографические ссылки.

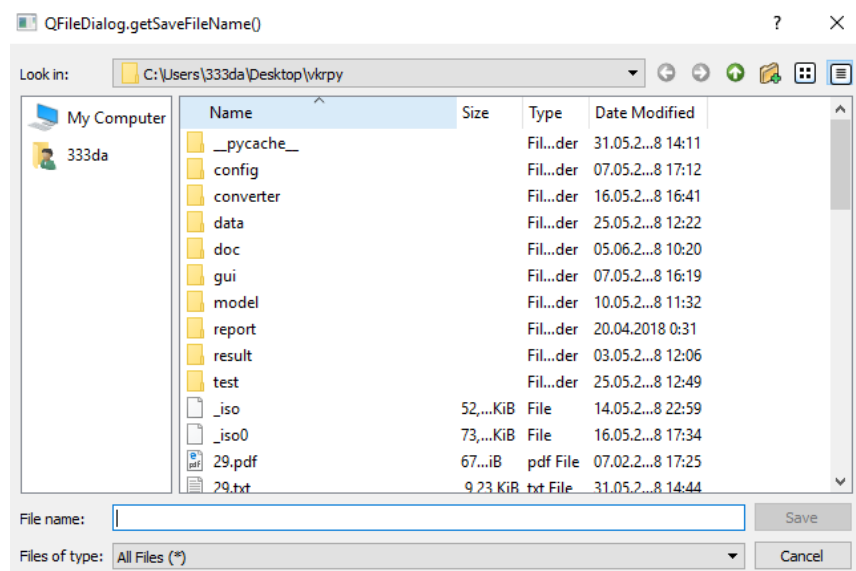
Выпадающий список дает возможность выбора формата для извлеченных метаданных *txt*, *iso19115v2*, *fgdc*.

Кнопка «*Extract*» - запускает процесс извлечения метаданных.

Кнопка «*Save*» - вызывает диалоговое окно Рис. 7, где пользователь выбирает место сохранения метаданных.



“Рис. 6 Открытие PDF файла”



“Рис. 7 Сохранения метаданных в файл”

2 Info (Рис. 8)

☒ Generate random UUID

UUID

Origin

Title

Date begin

Date end

Abstract

Supplemental

Purpose

Access

“Рис. 8 «Info»”

Чекбокс «*Generate random UUID*» - генерировать случайный идентификатор.

Поле «*UUID*» - уникальный идентификатор.

Поле «*Origin*» - Название организации или отдельного лица, которые разработали набор данных

Поле «*Title*» - Заголовок

Поле «*Date begin*» - Дата публикации.

Поле «*Date end*» - Дата окончания события

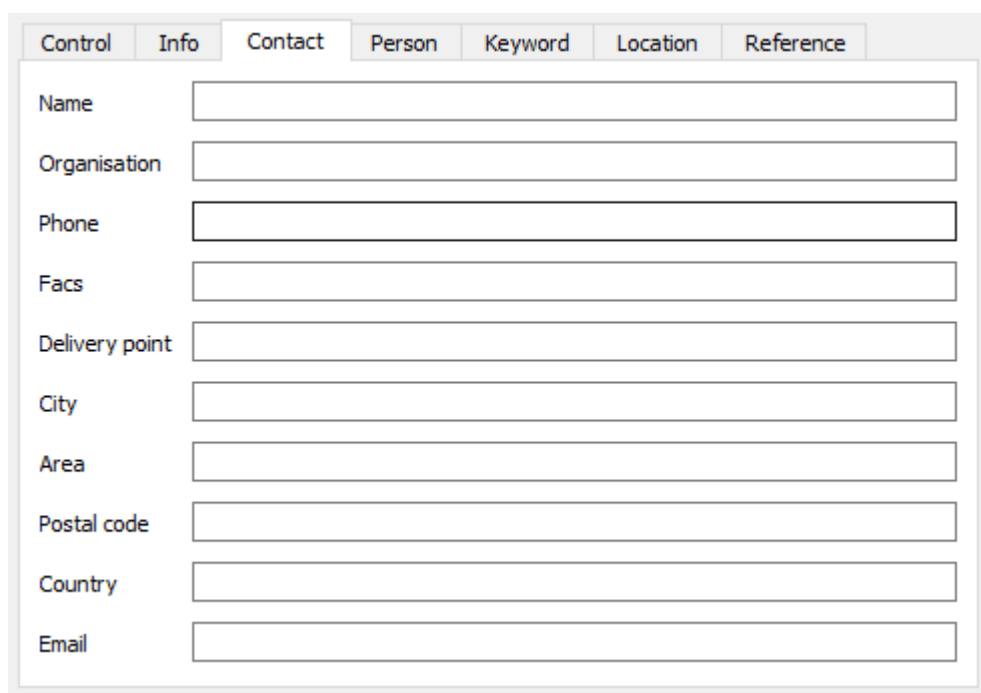
Поле «*Abstract*» - Краткое описательное данных

Поле «*Supplemental*» - Другая описательная информация о наборе данных

Поле «*Purpose*» - Краткое изложение целей, с которыми был разработан набор данных

Поле «*Access*» - Права доступа.

3 Contact (Рис. 9)



Control	Info	Contact	Person	Keyword	Location	Reference
Name		<input type="text"/>				
Organisation		<input type="text"/>				
Phone		<input type="text"/>				
Facs		<input type="text"/>				
Delivery point		<input type="text"/>				
City		<input type="text"/>				
Area		<input type="text"/>				
Postal code		<input type="text"/>				
Country		<input type="text"/>				
Email		<input type="text"/>				

“Рис. 9 «Contact»”

Поле «*Name*» - Имя человека

Поле «*Organization*» - Название организации

Поле «*Phone*» - Телефон

Поле «*Facs*» - Факс

Поле «*Delivery point*» - Адрес

Поле «*City*» - Город

Поле «*Area*» - Область, край

Поле «*Postal code*» - Почтовый или другой индекс

Поле «*Country*» - Страна

Поле «*Email*» - Адрес электронной почты

4 Person (Рис. 10)

Add	
1	Delete
Name	<input type="text"/>
Organisation	<input type="text"/>
Phone	<input type="text"/>
Facs	<input type="text"/>
Delivery point	<input type="text"/>
City	<input type="text"/>
Area	<input type="text"/>
Postal code	<input type="text"/>
Country	<input type="text"/>
Email	<input type="text"/>

“Рис. 10 «Person»”

Кнопка «Add» - добавляет новую запись.

Кнопка «Delete» - удаляет запись.

Поле «Name» - Имя человека

Поле «Organization» - Название организации

Поле «Phone» - Телефон

Поле «Facs» - Факс

Поле «Delivery point» - Адрес

Поле «City» - Город

Поле «Area» - Область, край

Поле «Postal code» - Почтовый или другой индекс

Поле «Country» - Страна

Поле «Email» - Адрес электронной почты

5 Keyword (Рис. 11)

Control Info Contact Person **Keyword** Location Reference

Key words

Add

1	Delete
---	--------

Name

Type

Locations

Add

1	Delete
---	--------

Name

Type

“Рис. 11 «Keyword»”

Кнопка «Add» - Добавляет новую запись.

Кнопка «Delete» - Удаляет запись.

Поле «Name» - Ключевое слово / словосочетания.

Поле «Type» - Тип ключевого слова.

6 Location (Рис. 12)

Add	
1	Delete
Name	<input type="text"/>
West	<input type="text"/>
East	<input type="text"/>
North	<input type="text"/>
South	<input type="text"/>

“Рис. 12 «Location»”

Кнопка «Add» - Добавляет новую запись.

Кнопка «Delete» - Удаляет запись.

Поле «Name» - Название локации.

Поле «West» - Долгота крайней западной точки.

Поле «East» - Долгота крайней восточной точки.

Поле «North» - Широта крайней северной точки.

Поле «South» - Широта крайней южной точки.

7 Reference (Рис. 13)

Control Info Contact Person Keyword Location Reference

Add

1	Delete
---	--------

Origin

Date

Title

Link

“Рис. 13 «Reference»”

Кнопка «Add» - Добавляет новую запись.

Кнопка «Delete» - Удаляет запись.

Поле «Origin» - Авторы.

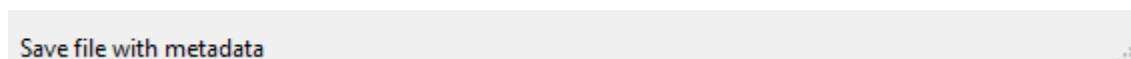
Поле «Date» - Год публикации.

Поле «Title» - Заголовок.

Поле «Link» - Ссылка на данные.

Статус бар (Рис. 14)

Выводит информацию от приложения.

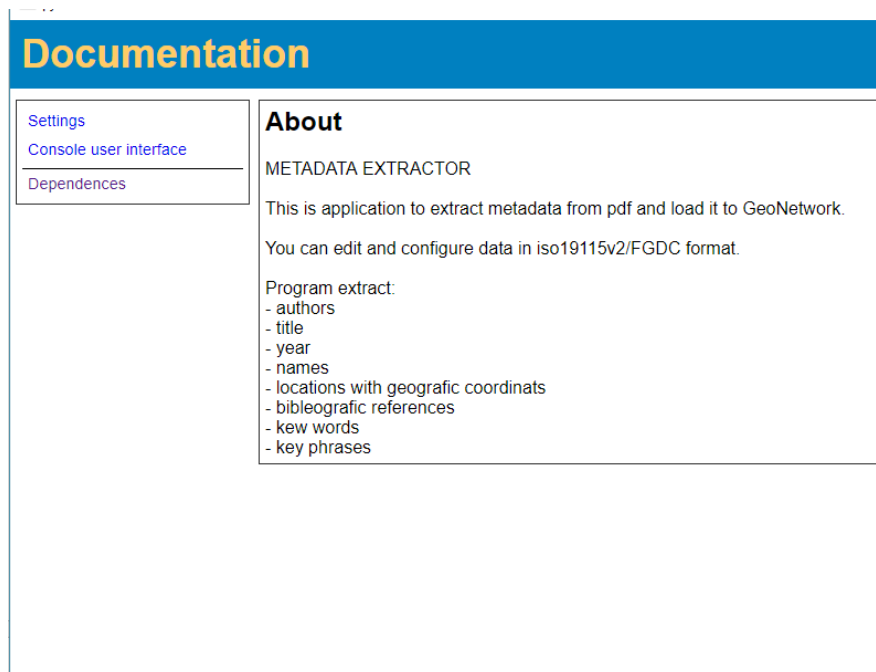


“Рис. 14 «Статус бар»”

Помощь

В меню “? -> Help” или “F1” (Рис. 15).

Здесь можно найти информацию о приложении.

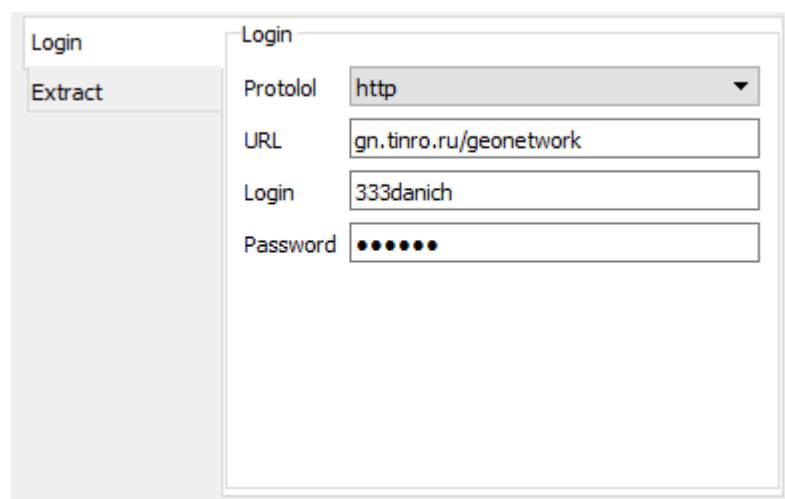


“Рис. 15 «Help»”

Настройки

В меню “? -> *Settings*” или “F3” (Рис. 16-17).

Настройки разбиты на вкладки.



“Рис. 16 «Settings Login»”

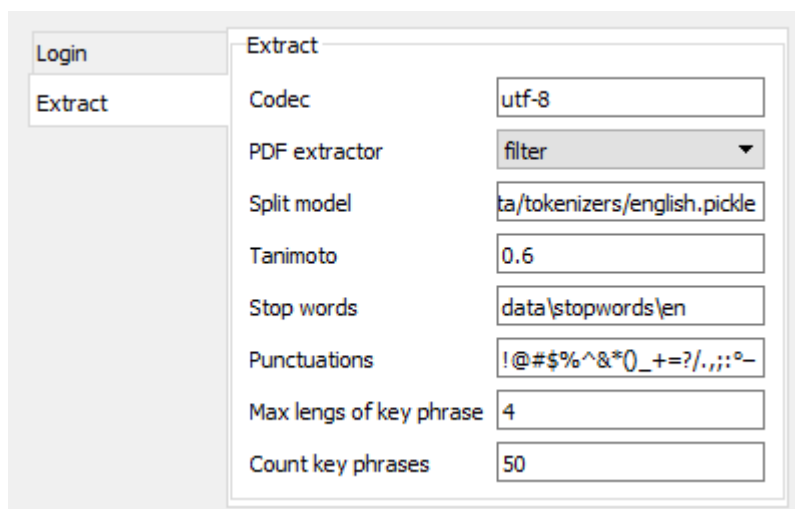
Описание вкладки «*Login*» (Рис. 16)

Выпадающий список «*Protocol*» – http / https.

Поле «*URL*» - адрес сервера GeoNetwork.

Поле «*Login*» - имя пользователя GeoNetwork.

Поле «*Password*» - пароль пользователя GeoNetwork.



“Рис. 17 «Settings Extract»”

Описание вкладки «Extract» (Рис. 17)

Поле «*Codec*» - Кодек входного файла.

Выпадающий список «*PDF extractor*» – «filter» / «text». «text» – быстрее.
«filter» – удаляет ненужные элементы (таблицы, графики, ...)

Поле «*Split model*» - Путь к модели для разделения текста на предложения.

Поле «*Tanimoto*» - Коэффициент для неточного сравнения строк.

Поле «*Stop words*» - Путь к списку со стоп словами.

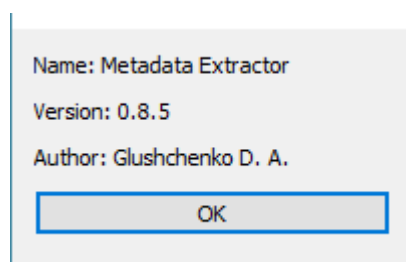
Поле «*Punctuation*» - Символы пунктуации.

Поле «*Max length of key phrase*» - Максимальная длина для ключевой фразы.

Поле «*Count of key phrases*» - Максимальное количество ключевых фраз.

О программе

Просто название программы, её версия и автор (Рис. 18)



“Рис. 18 «About»”

Работа через командную строку

Для работы через командную строку запускается скрипт *main.py* с ключами.

Ключи приведены в таблице 4.

Команда	Описание
--type -t	Тип выходного файла <i>txt</i> , <i>iso19115</i> или <i>fgdc</i>
--input_file -in	Название входного файла
--output_file -out	Название выходного файла
--metaTitle	Извлечение заголовка
--metaContent	Извлечение оглавления
--metaName	Извлечение авторов
--metaLocation	Извлечение локаций
--metaKeyWord	Извлечение ключевых слов и словосочетаний
--metaRef	Извлечение библиографических ссылок
--metaOrg	Извлечение организаций
--metaMisk	Извлечение разного
--metaAll	Извлечение всех метаданных
--formatPdf	Метод извлечений текста из pdf файла: <i>filter</i> – с фильтрацией, <i>text</i> – без фильтрации
--start	Начальная страница для извлечения метаописаний
--end	Конечная страница для извлечения метаописаний
--help	Помощь

“Табл. 4 Команды”

Пример: `main.py --metaAll --start 10 --end 20 --out out.txt -t txt -in in.pdf`

Из файла «*in.pdf*» со страницы 10 по страницу 20 извлекаются все метаданные и сохраняются в файл «*out.txt*» в формате «*txt*».

Другие настройки

В файле «*config/dictionary.py*» - хранится словарь с географическими объектами и их координатами. Его можно дополнять вручную.

Краткая инструкция по использованию программы

1. Запускаем «mainGUI.py»
2. На вкладке «Control» нажимаем кнопку «Open».
3. На вкладке «Control» отмечаем чекбоксы с нужными данными.
4. На вкладке «Control» из всплывающего списка выбираем нужный формат.
5. На вкладке «Control» нажимаем кнопку «Extract», ждём окончания процесса извлечения.
6. После завершения процесса извлечения на вкладках «Info», «Contact», «Person», «Keyword», «Location» и «Reference» можно подправить и дополнить данные.
7. Нажать на кнопку «Save» для сохранения данных на жесткий диск или нажать на кнопку «Load» в баре для загрузки на сервер GeoNetwork. (В настройках должны быть заполнены данные о сервере GN и данных пользователя.)