

Спецификация выходного JSON СмартПарсера, версия 2.0

Вводная информация	1
Описание JSON формата	3
Общая структура	3
Поле document (реквизиты входного документа)	3
Поле persons (все декларанты документа)	4
Поле ФИО декларанта person/name_raw	5
Поле person/department (отдел или подведомственная организация, где работает декларант)	5
Поле person/role (должность декларанта)	5
Поле адреса внутри документа (document_position)	6
Порядковый номер персоны (person_index)	6
Родственники (тип <relative_type>)	6
Сумма годового дохода (incomes)	6
Транспортные средства (поле vehicles)	7
Недвижимость (поле real_estates)	8
Сведения о расходах (поле spendings)	9
Авторизация в API	10
1) Basic HTTP Authorization	10
2) Token Authorization	11
API для валидации JSON файлов	11
Список валидаторов JSON	12

Вводная информация

Декларатор - единая база сведений о доходах и имуществе публичных должностных лиц (антикоррупционные декларации) и аналогичных сведений в отношении кандидатов на выборные должности (предвыборные декларации). Оба вида деклараций публикуются разными органами в разных форматах (Excel, Word, PDF, HTML); стандартизация хотя и существует на нормативном уровне, но остается недостаточной. Для качественного использования публикуемых данных требуется их приведение в единый машиночитаемый формат и дальнейшая обработка (связывание отдельных записей в цепочку, относящуюся к одному должностному лицу). Для

краткости мы будем называть **декларантами или персонами** всех должностных лиц, которые публикуют сведения о своих доходах.

Первая часть этого процесса - парсинг исходных файлов. Данные парсятся в формат json, описанные далее, который является промежуточным звеном при обработке данных. Импортируемые данные образуют дальнейшую обработку (связывание разделов), не описанные в настоящем руководстве. Разработка парсеров ведется, в том числе в открытом режиме в [репозитории Трансперенси Интернешнл - Россия на Гитхабе](#). На текущий день парсер называется "СмартПарсер" или smart_parser. Если вам потребуется консультация по этому документу, пишите нам на mail@declarator.org.

В сведениях о доходах и имуществе, публикуемых в соответствии с действующим антикоррупционным законодательством, указывается следующая информация:

- 1) Фамилия, имя, отчество декларанта;
- 2) Сумма годового дохода (руб.)
- 3) Транспортные средства;
- 4) Недвижимость;
- 5) Должность декларанта;
- 6) Департамент, в котором работает декларант (или подведомственная организация)
- 7) Доход, транспортные средства и недвижимость для супруг и несовершеннолетних детей без указания ФИО

Пример входной таблицы [1]:

Наименование организации	Фамилия, имя, отчество (Руководитель, супруг (а) дети)	Должность	Общая сумма декларированного дохода за 2012 г. (руб.)	Перечень объектов недвижимого имущества, принадлежащих на праве собственности			Перечень транспортных средств, принадлежащих на праве собственности (вид, марка)
				Виды объектов недвижимости	Площадь (кв.м.)	Страна расположения	
ФГБОУ ВПО «Дагестанский государственный педагогический университет»	Абдулаев Магомед Имранович	И.о. ректора	2 202 496,17р.	Земельный участок (индивидуальная собственность)	700	Российская Федерация	Mercedes Benz GLK-350 (индивидуальная)
				Квартира (индивидуальная собственность)	157	Российская Федерация	
				Земельный участок (индивидуальная собственность)	689	Российская Федерация	
	Супруг (супруга)		996 062,33р.	Квартира (индивидуальная собственность)	49,3	Российская Федерация	нет
				нет	нет	нет	нет
	Несовершеннолетняя дочь		0,00р.	нет	нет	нет	нет

Схема JSON файла доступна по ссылке:

<https://declarator.org/static/api/import-schema.json>

ВНИМАНИЕ: Файл не будет импортирован, если он не проходит валидацию схемы!

Описание JSON формата

Общая структура

Один json-файл имеет следующую структуру:

```
{
  "persons": [массив всех декларантов, найденных в документе ],
  "document": {описание документа, который был распарсен}
}
```

Поле document (реквизиты входного документа)

В поле document идёт описание документа, который был обработан парсером в процессе парсинга. В нём должны быть следующие поля:

1. documentfile_id - id файла документа из базы декларатора.
Команда SmartParser получает файлы распакованные и переименованные в формате:

- a. documentfile_id.ext если исходный файл был документов
 - b. documentfile_id/filename.ext если исходный файл в БД был архивом
2. archive_file - если исходный файл был архивом (случай b), то в этом поле должен быть указан файл (filename.ext), который был обработан парсером.
3. sheet_number - для Excel документов, рекомендуется разбивать JSON по листам, указывая в этом поле номер листа, данные из которого содержатся в текущем JSON файле.
4. sheet_title - для Excel документов, полный заголовок листа, обычно печатается в верхних строках документа (например, "Сведения о доходах, об имуществе и обязательствах имущественного характера федеральных государственных гражданских служащих Министерства сельского хозяйства Российской Федерации и членов их семей за отчетный период с 1 января 2012 г. по 31 декабря 2012 г. Департамент регулирования агропродовольственного рынка, рыболовства, пищевой и перерабатывающей промышленности")
5. year - заполняется, если известен год декларируемых доходов

Поле persons (все декларанты документа)

В поле persons идёт массив всех персон, формат которых подробно описан ниже. Основные поля каждого элемента этого массива:

```
{
  "person": {
    'name_raw': "string",
    'department': "string",
    'role': "string",
  },
  "document_position": "string",
  "person_index": "number"
  "incomes": [array],
  "real_estates": [array],
  "vehicles": [array],
  "spendings": [array]
}
```

Подробные описания всех объектов есть в JSON Schema.

Отдельно, требования по данным указаны ниже, вместе с описанием методов парсинга.

Поле ФИО декларанта person/name_raw

Встречается два основных способа написания в декларациях:

- 1) Фамилия Имя Отчество (Иванов Иван Иванович)
- 2) Фамилия И. О. (Иванов И. И.)

Обычно эта информация записана в одной ячейке таблицы, если отчества и имена записаны в разных ячейках таблицы, нужно собрать ФИО в одну строку в одном из двух указанных форматах. Если ФИО содержит больше трех слов, так его и оставлять - это разрешено.

Поле person/department (отдел или подведомственная организация, где работает декларант)

Отдел, в котором работает декларант, может находиться в отдельном столбце (как в примере[1]), а может стоять выше по таблице, когда декларанты идут группами, где каждая группа декларантов работает в одном отделе:

Пример входной таблицы [2]

Фамилия, имя, отчество	Должность	Общая сумма декларированного годового дохода за 2012 год (руб.)	Перечень объектов недвижимого имущества, принадлежащих на праве собственности или находящихся в пользовании			Перечень транспортных средств, принадлежащих на праве собственности (вкл. марка)	
			Вид объектов недвижимости	Площадь (кв.м.)	Страна происхождения		
Руководство территориальных органов и образовательных учреждений ФСИН России							
Заев Ю.Ю.	Начальник УФСИН России по Республике Адыгея (Адыгея)	795 449,40	земельный участок (собственность, общая долевая)	490,0	Россия	автомобиль легковой Киа Сонет	
			квартира (пользование)	48,0	Россия		
			гаражный бокс (собственность)	22,4	Россия		
Супруга		84 000,00	квартира (собственность)	63,3	Россия	-	
			квартира (пользование)	48,0	Россия		
Чернов А.Г.	Заместитель начальника УФСИН России по Республике Адыгея (Адыгея)	708 732,00	земельный участок (собственность)	741,0	Россия	автомобиль легковой Тойота Королла	
			квартира (собственность)	61,1	Россия		
			квартира (пользование)	53,4	Россия		
Руководство нетерриториальных органов и образовательных учреждений ФСИН России							
Чиназиев А.В.	Заместитель начальника УФСИН России по Республике Адыгея (Адыгея)	510 551,22	квартира (собственность, 1/4 доли)	61,2	Россия	автомобиль легковой Опель Вектра	

В последнем случае у всех декларантов одной группы будет одинаково заполнено поле person/department. Поле person/department может отсутствовать, тогда считается, что все декларанты работают в головной организации, которая опубликовала этот документ.

Поле person/role (должность декларанта)

В поле person/role записывается должность, на которой работает декларант. Обычно эта должность идет в отдельном столбце (см. примеры выше).

Поле адреса внутри документа (document_position)

Для будущего анализа результатов парсинга нужно указать место в документе, где именно была обнаружена данная секция.

Для Excel - это адрес ячейки (например, "A:12")

Для Doc или PDF - это номер страницы документа.

Порядковый номер персоны (person_index)

Иногда в документе перед каждой персоной в отдельном столбце указан её порядковый номер в данном документе. В таких случаях этот номер надо указать в данном поле. Пример (первый столбец):

3	Краевой Сергей Александрович	Заместитель Министра	Земельный участок	индивидуальная	1833.0	РОССИЯ				Автомобиль ВАЗ (Бронто) 212140	15 063 446,77	
										Автомобиль. Друг		

ВНИМАНИЕ

Иногда порядковый номер указывается для департамента и объединяет несколько персон. В таких случаях это поле не нужно заполнять.

Родственники (тип <relative_type>)

Все объекты владения (доход, транспортное средство или недвижимость) содержат поля **relative** с допустимыми значениями: "Супруг(а)", "Ребенок" или null. Это поле обозначает, к кому относится данная запись, например, что машина супруги. Значение null означает, что эта запись относится к основному декларанту. Везде ниже мы будем писать, что поле relative принадлежит типу <relative_type> (может иметь указанные выше три значения: "Супруг(а)", "Ребенок" или null).

Сумма годового дохода (incomes)

Сумма годового дохода обычно указывается в отдельной ячейке. Почти всегда есть доход основного декларанта (отсутствие этого поля может случиться не по нашей вине). Кроме этого, в декларациях записана информация о доходах супруг(ов). Для каждого дохода вносится следующая запись:

```
{
  "size": number, ( float или integer)
  "size_raw": string
  "relative": <relative_type>
```

```
},
```

Поле **size** - это доход чиновника в рублях, иногда они пишут с точностью до копейки, тогда возникает не целочисленное, а число с плавающей точкой - это разрешено.

Поле **size_raw** - исходное сырое строковое значение ячейки с суммой дохода.

Во входных таблицах доход чаще всего встречается в виде числа с плавающей точкой, иногда с “руб.” или “р.” на конце, что нужно отбросить, так как по JSON схеме поле size должно быть number.

Внимание! Иногда сумма обозначается в тысячах рублей, что может быть указано как после цифры, так и в заголовке столбца. В таком случае приведенное в исходном файле число надо умножить на 1000. Существует еще формат “11 рублей 50 копеек”, который надо преобразовать к флоату (11.50).

Также иногда в декларациях встречается комментарий к источнику дохода, обычно указываемый рядом с суммой дохода в скобках, например, “часть дохода получена от продажи дома”. Эти строки надо внести в поле `"/persons/incomes/comment"`.

В случае, когда доходы не указаны совсем, следует оставить пустым список в JSON `"/persons/incomes"`.

Транспортные средства (поле vehicles)

Транспортные средства обычно указываются в отдельных столбцах, и имеют следующие поля:

```
{
  "relative": <relative_type>
  "text" : string,
  "type_raw": string,
  "manufacture_year": number,
  "own_type_raw":string
}
```

Поле **relative** - обязательно поле, см. выше список констант, которые могут тут стоять.

Поле **text** - обязательное поле, которое обычно содержит описание одного транспортного средства. Пример записи о транспортном средстве:

```
"vehicles": [
  {
    "text": "а/м легковой Тойота Ланд Крузер Прадо",
    "relative": null
  }
]
```

Может быть и несколько транспортных средств в одной ячейке:

```
"vehicles": [
  {
    "text": "а/м легковой Ауди; Квадроцикл Ямаха",
    "relative": "Супруг(а)"
  }
]
```

Если описание транспортных средств разделено на разные колонки, тогда их обязательно надо слить (конкатинировать через пробел) в поле text, но и оставить в отдельных полях: .

- 1) Поле type_raw - вид транспортного средства - заполняется, если вид транспортного средства был указан в отдельной ячейке, например: "Автомобиль", "А/м", "Грузовой транспорт", "Лодка", "Снегоход"
- 2) Поле "manufacture_year" - год выпуска -указывается, если был указан в отдельной графе
- 3) Поле own_type_raw - тип владения заполняется сырым строковым значением, если был указан в отдельной ячейке.

Недвижимость (поле real_estates)

В сведениях о доходах обычно **отдельно** указываются объекты недвижимости в **собственности** и **находящиеся в пользовании**. Как правило, в исходной таблицы они располагаются в отдельных группах столбцов: левее группа из 4 столбцов, касающаяся недвижимости в собственности, правее 3 столбца о недвижимости в пользовании.

Фамилия и инициалы лица, чьи сведения размещаются	Должность	Объекты недвижимости, находящиеся в собственности				Объекты недвижимости, находящиеся в пользовании			Транспортные средства (вид, марка)	Декларированный годовой доход <1> (руб.)	Сведения об источниках получения средств, за счет которых совершена сделка <2> (вид приобретенного имущества, источник)
		вид объекта	вид собственности	площадь (кв. м)	страна расположения	вид объекта	площадь (кв. м)	страна расположения			
Елин Е.И.	Заместитель Министра	Квартира	Индивидуальная	127.40	Россия	Квартира	183.00	Россия	Автомобиль легковой Мерседес	3575166.70	-
		Квартира	Общая долевая (1/2)	85.00	Россия						

В некоторых случаях, оба типа недвижимости объединены в один столбец, а режим использования уточняется в скобках. Один объект недвижимости описывается следующими полями:

```
{
  "own_type_by_column": "В собственности" | "В пользовании"
  "own_type_raw": string,
  "relative": <relative_type>
  "type_raw": string,
  "square": float,
  "square_raw" : string,
  "country_raw": string
  "text": string
}
```

Например:

```
{
  "own_type_by_column": "В собственности",
  "own_type_raw": "Индивидуальная",
  "type_raw": "Земельный участок",
  "text": "Земельный участок",
  "square_raw": "1500.00",
  "square": 1500.00,
```



```

"country_raw": "Россия",
"relative": null
}

```

Для того, чтобы отделить эти два типа используется поле **own_type_by_column** (личной собственности соответствует значение “В собственности”, и для пользования - “В пользовании”).

Поле **own_type_raw** заполняется, когда присутствуют отдельный столбец с уточнением типа владения. Столбец обычно называется “вид собственности”, примеры значений такие:

1. “индивидуальная”,
 2. “долевая собственность 1/2”
 3. “общая собственность”
 4. “совместная собственность”
- и пр..

Поле **type_raw** содержит тип недвижимости (квартира, земельный участок, дача и пр.). В этом же поле содержится иногда указание доли владения.

Поле **square_raw** содержит площадь недвижимости в метрах квадратных. Поле **square** содержит преобразованное во float значение поля **square_raw**.

Поле **country_raw** содержит страну, в которой находится данный объект недвижимости.

Иногда **type_raw** и **own_type_raw** записаны во входной таблице в одной ячейке, в таких случаях, содержимое этой ячейки пишется в поле **text**, а поля **type_raw** и **own_type_raw** не заполняются.

Фамилия, имя, отчество	Должность	Общая сумма дохода за 2011 г. (руб.)	Перечень объектов недвижимости, принадлежащих на праве собственности			Перечень объектов недвижимости, находящихся в пользовании			Движимое имущество
			Вид объектов недвижимости, вид собственности	Площадь (кв.м.)	Страна <u>расположения</u>	Вид объектов недвижимости	Площадь (кв.м.)	Страна <u>расположения</u>	
Попов Николай Валерьянович	Статс-секретарь – заместитель федерального Министра	15863749	садовый участок (индивидуальная)	600	Россия	-	-	-	а/м Volvo XC 70

Если же **type_raw** и **own_type_raw** разделены, тогда в поле **text** пишется (дублируется) значение поля **type_raw**. Поле **text** является обязательным.

Недвижимость без разделения на колонки

Встречаются документы, где недвижимость указана в одной колонке, в одной строке включены сразу все данные - площадь, вид и тип владения. В таких случаях эту строку надо заносить в поле **text**.

Бывает, что в одной колонке вместе указывается вид недвижимости и тип собственности, которые также надо заносить в поле **text**, а площадь и страну заносить в соответствующие поля - **square_raw** и **country_raw**.

Пример:

Департамент управления делами

ФИО	Должность	Годовой доход	Перечень объектов недвижимого имущества, принадлежащего государственному служащему, его супруге (супругу) и несовершеннолетним детям на праве собственности или находящихся в их пользовании	Площадь	Страна расположения имущества	Перечень транспортных средств, с указанием вида и марки принадлежащих на праве собственности
Абрамова Елена Геннадьевна	заместитель начальника отдела	827 730,00	квартира (долевая собственность, 1/3)	63,2	Россия	—
супруг	—	1 403 037,00	квартира (долевая собственность, 1/2) квартира (безвозмездное пользование)	24,5 63,2	Россия	а/м Ford Focus
ребенок	—	—	квартира (безвозмездное пользование)	63,2	Россия	—
Баазов Руслан Баазович	начальник отдела	666 196,61	квартира (безвозмездное пользование)	83	Россия	—
			земельный участок (собственность)	500		

Иногда даже отсутствуют площадь и тип владения, есть только описание объекта.

Пример:

№ п/п	Фамилия и инициалы лица, чьи сведения размещаются	Должность	Объекты недвижимости, находящиеся в собственности	Объекты недвижимости, находящиеся в пользовании	Транспортные средства (вид, марка)	Декларированный годовой доход ¹ (руб.)	Сведения об источниках получения средств, за счет которых совершена сделка ² (вид приобретенного имущества, источники)
1.	Акимов В.А.	Начальник ФГБУ ВНИИ ГОЧС (ФЦ)	квартира, машиноместо	-	легковая, БМВ X 6	8423952,16 (с учетом продажи имущества)	-
2.	Дурнев Р.А.	Заместитель начальника	квартира	-	легковая, Пежо 4007	5184052,09	-

В таких случаях содержимое этой одной колонки целиком должно быть записано в поле **text**, но если колонки две (как в примере выше), то по ним должен быть определено и записано поле **own_type_by_column**.

Сведения о расходах (поле `spendings`)

Иногда, в документах указываются расходы, сумма которых превышает доход за три года, в таком случае заполняется поле `"/persons/spendings"` в виде массива строк всех расходов персоны.

Авторизация в API

Используются учётные данные пользователя Декларатора (обращайтесь к нам за ними, если у вас их ещё нет)

Авторизоваться для обращений к API, можно двумя методами (второй метод предпочтительнее):

1) Basic HTTP Authorization

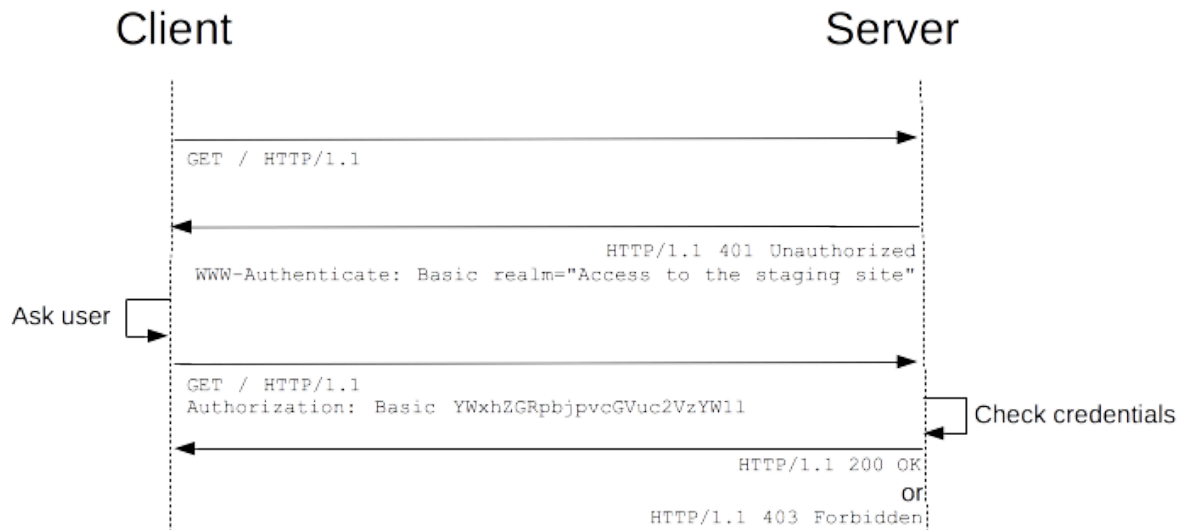
Пример использования curl для базовой HTTP авторизации:

```
curl --user login:password -X POST https://declarator.org/api/unknown\_entry/
```

Ваш скрипт должен уметь использовать [HTTP-заголовок Authorization](#), например:

```
Authorization: Basic QWxhZGRpbjpPcGVuU2VzYW11
```

Процесс авторизации происходит по стандартному протоколу:



2) Token Authorization

В процессе используется токен авторизации, который надо вставить в HTTP-заголовок обращения к API.

Для получения токена нужно сделать POST запрос на URL:

<https://declarator.org/api/api-token-auth/>

с JSON данными вида:

```
{
  "username": "ваш логин",
  "password": "ваш пароль"
}
```

В ответ вы получите JSON объект, где в поле “token” будет содержаться 40 символов, например: 9944b09199c62bcf9418ad846dd0e4bbdfc6ee4b. Это и есть ваш токен!

Для последующей авторизации к каждому HTTP запросу добавляется HTTP-заголовок вида:

```
Authorization: Token 9944b09199c62bcf9418ad846dd0e4bbdfc6ee4b
```

(где вместо 9944b09199c62bcf9418ad846dd0e4bbdfc6ee4b вы подставляете свой токен)

Пример использования curl с авторизацией при помощи токена:

```
curl -X POST http://declarator.org/api/unknown_entry/ -H 'Authorization: Token 9944b09199c62bcf9418ad846dd0e4bbdfc6ee4b'
```

API для валидации JSON файлов

В качестве проверки валидности JSON данных, которые получились после парсинга, можно воспользоваться API: <https://declarator.org/api/jsonfile/validate/>

Если ([авторизовавшись](#)) сделать HTTP POST на этот адрес с содержимым JSON файла в данных POST запроса, то в ответ можно получить список из невалидных записей, с комментарием в поле 'errors' - почему и какие данные не подходят для импорта.

Очень часто этот способ позволяет определить неверные данные, полученные в процессе парсинга.

Например, слишком маленький размер дохода персоны (size) можно получить, если неверно расшифровать float из экселя: 123.455 может означать "123 тысячи 455 рубля", а не "123 рубля и 45.5 копейки", как иногда случается. Такие проблемы предлагается исправлять непосредственно в процессе парсинга данных, не забывая при этом записывать исходные сырые значения в поле (size_raw).

Или же ФИО может содержать что-то недопустимое для импорта, например, слово "дочь". В таком случае можно сделать вывод, что парсер ошибочно записал родственника как отдельную персону.

В единичных случаях можно изменить исходный файл (excel или word), но для этого вам надо связаться с нами, чтобы мы не потеряли эти изменения и учитывали их в будущем.

Список валидаторов JSON

На каждое поле действует требование JSON schema, кроме этого ещё есть ряд дополнительных требований. Ниже приведён текст возможной ошибки (с комментарием):

1. RealEstate must have 'text' (or 'name') value
2. RealEstate text value is too long (>255 symbols)
3. Income must have size_raw
4. Unparsable float in income size
5. Too big income size / Too small income size (корректные значения от 100 до 10000000000)

6. "Cannot parse" on own_type/type/other dict value