



Протокол встреч 04.04.2022

Оглавление

1. Общая информация.....	2
2. Протоколы встреч.....	3
2.1. Общие технические требования.....	3
2.1.1. Обсуждение архитектуры взаимодействия и требований к интеграции внутренних систем...	3
2.1.2. Обсуждение функциональных требований Окулус-1 и ИС Вепрь с участием представителей СМК.	5
Приложение 1. Схема процессов взаимодействия сервисов ИС МИР.....	7



1. Общая информация

Цель встреч – экспертная оценка ресурсов и возможностей реализации проектов документов:

- Технические требования на создание автоматизированной системы выявления признаков нарушений законодательства Российской Федерации в изображениях и видеоматериалах информационной системы Мониторинг информационных ресурсов (шифр «ОКУЛУС-1»)
- Техническое задание на выполнение работ первого этапа создания информационной системы раннего выявления угроз в информационной сфере и прогнозирования рисков их возникновения (ИС «ВЕПРЬ»)

Формат проведения

ВКС, доступ для участников по ссылке:

<https://vks.grfc.ru/client?conference=7714&pin=MzIxOA%3D%3D>

Состав участников:

№	Время	Организатор	ГРЧЦ (НТЦ)	ГРЧЦ (СМК)	Социальная Лаборатория
1	13:00-14:30	Лилия Минакова, Менеджер проекта ИС МИР ОКУЛУС-1, НТЦ	Федотов Александр, Руководитель НТЦ Волошина Мария, (Администрация) Лубянцеv Геннадий (менеджер проекта ИС ВЕПРЬ) Евгения (Менеджер проекта ИС МИР ОКУЛУС-1)		Наталья Тылевич, Генеральный директор
2	15:30-16:30	Лилия Минакова	Федотов Александр, Ковалев Владислав, зам. Начальника по внутренним проектам НТЦ Невзоров Андрей, Руководитель ML-обработки данных отдела внутренних проектов НТЦ Антон Корнеев (Интеграция и сопровождение)	Елена, Мария,?	Наталья Тылевич, Генеральный директор
3	16:30-17:30	Лубянцеv Геннадий, менеджер проекта ИС ВЕПРЬ	Невзоров Андрей, Руководитель ML-обработки данных отдела внутренних проектов НТЦ	Александр Спиридонов Елена, Мария	Наталья Тылевич, Генеральный директор



2. Протоколы встреч

2.1. Общие технические требования.

Повестка:

2.1.1. Обсуждение архитектуры взаимодействия и требований к интеграции внутренних систем.

№	Вопрос	Обсуждение	Задачи, ответственные, сроки
1	Какие ожидания в отношении модуля ОКУЛУС-1 и ВЕПРЬ есть на текущий момент? Как выглядит текущий инструментарий, что уже автоматизировано, откуда поступают данные?	Федотов А.: Ключевые блоки функционала ИС МИР: - Система сбора данных - Единая мультимодальная система обработки и анализа данных, в составе которой есть и будут развиваться инструменты анализа, релевантные типам контента: - текст - звук - видео и изображения Ключевая задача ОКУЛУС-1 – обеспечить инструментарий выявления контента, соответствующего перечню приоритетных для выявления нарушений (приложение 1 к ТЗ). Для проведения работ и фиксации критериев ПМИ могут быть предоставлены материалы в соответствии с задачами распознавания: 200 лиц/персон, присутствие которых на изображениях необходимо будет выявлять автоматически. - Список экстремистских организаций, утвержденный с примерами флагов и иной символики (также, порядка 200 организаций) - Наиболее узкое место – выявление сцен. Как сформировать наборы данных для обучения в этой части вопрос открытый. Особенность формирования выборки: Формирование парных примеров изображений, - просто поездов – примеры поездов с «защеперами» - пример просто людей на крыше, экскурсии по крышам – пример крыш с людьми, совершающими суицидальные попытки. - Примеры людей с оружием и тд. Более конкретная схема процессов взаимодействия сервисов в составе ИС МИР представлена в приложении 1.	1. Подготовка Перечня 200 персон. 2. Подготовка базы данных с фотографиями для реализации обучения алгоритма по выявлению этих персон. Ответственный: Федотов А., Минакова Л. Сотрудники СМК (вопрос для обсуждения за рамками встречи внутри НТЦ).
2	Есть ли описание точек REST-API? Какой перечень этих точек?	Корнеев А. Взаимодействие сервисов выстроено по протоколу S3 API, перечень точек может быть подготовлен и предоставлен. Основные системы администрирования: Kubernetes, Grafana, Kafka права администрирования и уровней доступа настраиваются в KeyCloack и не являются значимой задачей на текущем этапе. Логирование и управление задачами через очереди Rabbitmq. Не ожидается проблем в части интеграции системы в общую среду. На текущий момент есть трудности в переносе сервиса краулинга в среду Kubernetes из SWARM. *Тылевич Н. Функционал краулера был урезан в соответствии с требованиями и специально упакован в SWARM для передачи по проекту ИС МИР-1.	*Организовать коммуникацию между девопсами ГРЧЦ и СЛ по вопросам развертывания функционала краулера в среде Kubernetes. Ответственный: Тылевич Н. Срок (в порядке приоритетов по подготовке и запуску ТЗ этого года).



Какой перечень моделей? Кто выбирает их применение? Как должна осуществляться последовательность выполнения набора моделей? Что должно отображаться? Кому (какой группе пользователей)?	Тылевич Н.: Контекст вопроса: Перечень Приложения 1 является широким и подобрать корректные примеры для выделения подклассов нарушений не представляется возможным, т.к. визуальный ряд может не иметь существенных отличий на картинках (Пример экстремизма – сговор). Возможны следующие варианты решений: 1. Прорабатывать отдельные классификаторы для картинок и текстов, опираясь на базу реальных примеров, а текущий классификатор считать «общим видением». Вопрос ответственности за подготовку и согласование классификаторов, которые могут лечь в основу ПМИ – обоюдная ответственность. Команда СЛ готова подготовить минимальные наборы данных с примерами, чтобы на практике обсуждать примеры публикаций, их состав по тексту и картинкам, чтобы валидировать такой классификатор совместными усилиями со стороны СМК (ГРЧЦ). 2. Поскольку конечные пользователи со стороны СМК работают с публикациями в целом, учитывая набор параметров: Источник публикации/автор, текст публикации как вариант интерпретации, статистика по охвату и тд. Стоит рассмотреть следующие варианты решений: 2.1. Включение сведений о контексте изображения в структуру базы данных по картинкам и разработку алгоритмов анализа, учитывающих весовые коэффициенты в выявлении изображений с нарушениями (то есть внести соответствующие изменения в структуру базы данных по картинкам, по аналогии с РИС системами по хранению изображений типа DICOM). 2.2. Развитие архитектуры ИС в части функционала проектного подхода к управлению процессами анализа: Т.е. Внедрение инструментария формирования обучающей выборки и дообучения нейронных сетей в рабочие процессы и создание гибких классификаторов внутри соответствующих проектов, с более узкими классификаторами. Для подразделений, отслеживающих экстремизм – свои классификаторы и наборы данных для автоматизации разметки, для подразделений, осуществляющих работу по выявлению порно – свои, для наркотиков, курения – свои. В любом случае, необходимо понимать: На каких данных будет проводиться обучение На каких данных будет поводится приемка Если присутствует задача по интеграции с сервисом краулинга, то необходимо заблаговременно определить и настроить сбор данных по источникам, которые будут утверждены как выборка для порведения испытаний вместе с наборами поисковых запросов к этим источникам. Невзоров А.: Я понимаю задачу по ключевым словам для проработки классификатора. Изучим возможность проработки требований в этой части. Резюме: Текущий перечень нарушений, приоритетных для выявления необходимо дорабатывать и привести к конкретному перечню критериев оценки работоспособности системы, который ляжет в основу ПМИ, оценки объемов и пиритности работ в составе проекта.	Задача: Обсуждение возможных развилок по подходу к развитию ИС в целом внутри ГРЧЦ. Ответственный: Федотов А. Подготовка перечня ключевых слов для формирования классификаторов разметки данных и последующего обучения. Ответственные: Невзоров А. (на базе существующих подходов к обработке данных/опыте НТЦ и СМК) Тылевич Н. Подготовка выборки для валидации совместно с сотрудниками СМК и согласованию классификаторов. Сроки подготовки минимальных наборов данных: 12-13 апреля.
---	---	---



2.1.2. Обсуждение функциональных требований Окулус-1 и ИС Вепрь с участием представителей СМК.

№	Вопрос	Обсуждение	Задачи, ответственные, сроки
1	Какие ожидания в отношении модуля ОКУЛУС-1 и ВЕПРЬ есть на текущий момент? Как выглядит текущий инструментарий, что уже автоматизировано, откуда поступают данные?	Спиридонов А.: Мои ожидания таковы: нам нужен весь русскоязычный интернет, все публикации на любых источниках. Типы источников для СМК: - зарегистрированные сми (WEB) - незарегистрированные СМИ - соц.сети и каналы В нашем случае база источников не является стабильной как в случае, например, с протестной активностью, ресурсами, аффилированными с соответствующими группами влияния (к таким ресурсам можно приставить человека, это решается). Нас прежде всего интересуют новые источники в соц. Сетях. (телеграм-каналы, приватные чаты создаются зачастую буквально на несколько часов для обмена информацией. У них узкая аудитория, но нам надо их выявлять и блокировать. Поэтому нам нужен весь интернет и все источники. Критерии выявления публикаций в связи с этим для нас в приоритете именно статистические: увидели всплеск количества публикаций – значит идет волна. Донастройки по тематикам, типам источников, инструментарию анализа аудитории – это все в меньшем приоритете, на других этапах. Тылевич Н. Требования к расширению пула источников являются задачей сервиса мониторинга и не относятся к функционалу обработки изображению (ОКУЛУС-1) и выявлению инцидентов. Для проведения работ и успешной приемки необходимо определить критерии и ограничить выборку объема данных. Спиридонов А. Нет, это не интересно. Нужны все источники без ограничений. На приемке я буду сравнивать то, что мы нашли руками с тем, что найдет система и если будут пропуски – значит система не работает. Тылевич Н. При такой постановке задачи потребуются колоссальные мощности по инфраструктуре серверов, а также существенное расширение функционала краулинга в части настройки первичной базы источников (система может автоматически выявлять ссылки на новые источники внутри текста публикаций, но при таком естественном росте базы источников необходимо: - учитывать кластеризацию источников. С большой степенью вероятности, искомый источник входит в кластер(ы), конкретную связанную сеть распространения информации. То есть новый источник на этапе запуска для привлечения аудитории стремится образовывать связи (ссылку на свой ресурс как источник информации) на источник с БОЛЬШИМ объемом аудитории. Как только известный источник размещает информацию со ссылкой на новый, раскрываемый источник, возникает возможность подключения этого источника. В случае, если исходная база источников мониторинга узкая, есть вероятность, что кластеры источников будут охвачены достаточно быстро и прирост источников остановится, т. к. Не будет пересечений публикаций с новыми целевыми источниками. NB! Развитие базы источников и инструментов их администрирования	Провести внутреннее обсуждение конфликта ожиданий в части технических и функциональных требований. Рассмотреть возможные варианты решений. Ответственные: Федотов А. Минакова Л. Лубянец Г.



		– самостоятельная емкая задача, которая не является частью обсуждаемых ТЗ. Возможные варианты решений: - Определить и утвердить перечень источников для подключения в рамках проекта, результаты сбора данных по которому смогут стать выборкой для проведения ПМИ - Рассмотреть возможности более интенсивного наращивания базы источников за счет достижения политических договоренностей с vk, ok, поисковиками и счетчиками. NB! Базы данных по всем русскоязычным источникам всех типов не существует, невозможно оценить объем мониторинга по неограниченной выборке, скорость роста объемов, провести нагрузочные тестирования и тд. При такой постановке требований оценка и последующая приемка результатов работ не представляется возможной.	
--	--	--	--

Резюме:

1. Для проведения оценки стоимости реализации ТЗ ОКУЛУС-1 необходимо проведение работ по разработке классификатора и ПМИ. Текущий разброс имеет риски кратного увеличения объемов работ в процессе реализации.
2. Для проведения оценки стоимости реализации ТЗ ВЕПРЬ необходимо определить критерии приемки. Возможно, пересмотреть последовательности приоритетов в развитии функционала системы (сначала провести работы по развитию и настройке системы сбора данных, а потом уже переходить к работе над аналитическими инструментами по выявлению и идентификации рисков. Либо скорректировать ожидания внутреннего функционального заказчика и определить достижимые критерии приемки данного функционала с учетом общего уровня развития сервисов системы).



Приложение 1. Схема процессов взаимодействия сервисов ИС МИР



