



**Vilniaus  
universitetas**



---

# **Ataskaitinė informatikos krypties doktorantų konferencija 2022-03-24**

Rolandas Gricius (VU DMSTI doktorantas, Išmaniųjų technologijų tyrimų grupė)



**Darbo tema.**

Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose.

Recognising the contents in digitised structured documents.

**Darbo vadovas.**

Prof. dr. Igoris Belovas.

**Doktorantūros studijų laikotarpis.**

2021 m. spalio mėn. 1 d. – 2025 m. rugsėjo mėn. 30 d.

**Ataskaitinis laikotarpis.**

2021 m. spalio mėn. 1 d. – 2022 m. kovo mėn. 24 d.

# Visų studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė
I (2021/2022) Pirmas pusmetis	1	1		1 (L)			
II (2021/2022) Antras pusmetis	1						
II (2022/2023) Pirmas pusmetis	1						
II (2022/2023) Antras pusmetis	1		1 (T)				
III (2023/2024) Pirmas pusmetis							
III (2023/2024) Antras pusmetis					1 (CA WoS)		
IV (2024/2025) Pirmas pusmetis							
IV (2024/2025) Antras pusmetis			1 (T)		1 (CA WoS)		

# Ataskaitinių metų darbo planas ir jo vykdymo suvestinė

Egzaminai		Dalyvavimas konferencijose		Publikacijos	
Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta
Mašininis mokymasis	Išlaikyta: Mašininis mokymasis, 2022-02-17, pažymys 8.	-	Nuotoliniu būdu dalyvauta: 12-ojoje tarptautinėje konferencijoje „Duomenų analizės metodai programų sistemoms“, Druskininkai, Lietuva, 2021-12-03 pristatytas stendinis pranešimas.	-	-

## Visų mokslinių tyrimų ir disertacijos rengimo etapai

Darbo pavadinimas		Atlikimo terminai	Pastabos
1.	<b>Mokslinių tyrimų disertacijos tema apžvalga ir analizė:</b>	2021 m. spalio mėn. – 2022 m. kovo mėn.	Paruošta mokslinės literatūros apžvalga, kuri bus užbaigta pagal planą kitą pusmetį. Detalizuotas disertacijos tyrimo objektas.
	1.1. Analitinės apžvalgos atlikimas. 1.2. Disertacijos tyrimo objekto detalizavimas.		
	1.3. Analitinės apžvalgos užbaigimas. 1.4. Mokslinių problemų susietų su tyrimo objektu identifikavimas ir tyrimo tikslo suformavimas.	2022 m. balandžio mėn. – 2022 m. rugsėjo mėn.	
2.	Mokslinio tyrimo vykdymas:		
	<b>2.1. Tyrimo metodikos sudarymas</b>	2022 m. spalio mėn. – 2023 m. kovo mėn.	
	<b>2.2. Teorinis tyrimas</b>	2023 m. balandžio mėn. – 2023 m. rugsėjo mėn.	
	<b>2.3. Empirinis tyrimas</b>	2023 m. spalio mėn. – 2024 m. rugsėjo mėn.	
	<b>2.4. Gautų rezultatų analizė ir apibendrinimas</b>	2024 m. spalio mėn. – 2025 m. kovo mėn.	

## Visų mokslinių tyrimų ir disertacijos rengimo etapai

Darbo pavadinimas		Atlikimo terminai	Pastabos
3.	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų ir kt.) parengimas: 3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas. 3.2. Analitinės disertacijos dalies parengimas. 3.3. Teorinės disertacijos dalies parengimas. 3.4. Eksperimentinės disertacijos dalies parengimas. 3.5. Bendrųjų išvadų formulavimas.	2025 m. balandžio mėn. – 2025 m. rugsėjo mėn.	
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2025 m. birželio mėn.	
5.	Daktaro disertacijos gynimas	2025 m. rugsėjo mėn.	



# Tyrimo objektas, tikslas ir uždaviniai

„Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose“

- Dokumentai
- Struktūrizuoti dokumentai
- Suskaitmeninti struktūrizuoti dokumentai
- Turinio atpažinimas

---

# Dokumentai



- Dokumentai neberašomi ranka
- Spausdinami arba siunčiami el. paštu
- Dokumento duomenys suvedami į gavėjo sistemas



---

# Struktūrizuoti dokumentai



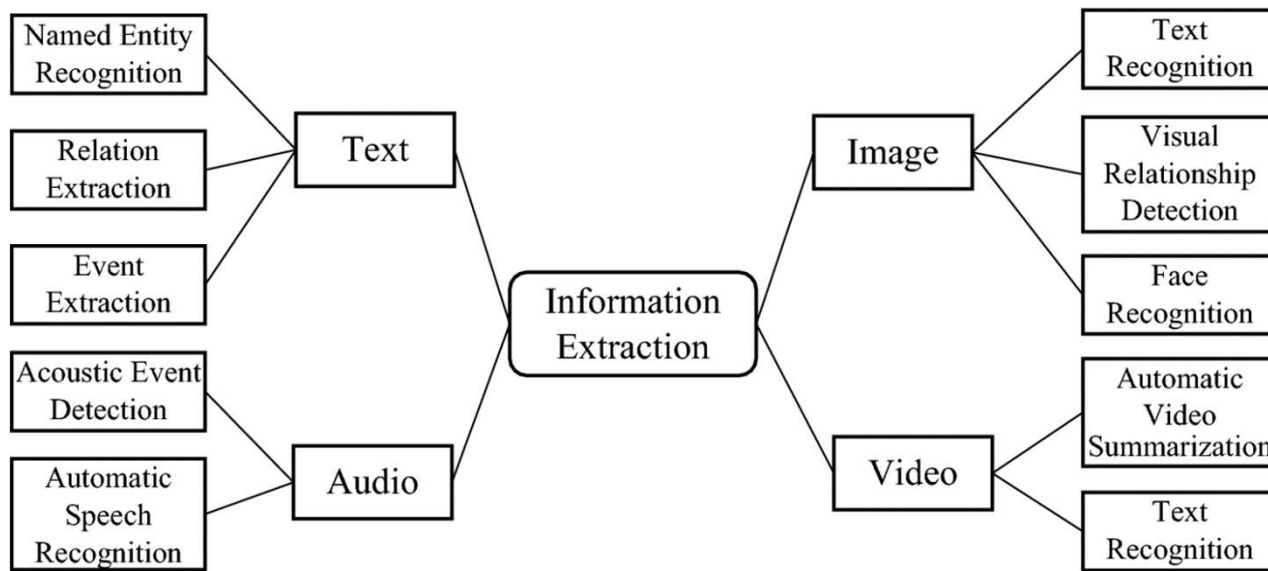
- **Griežtai struktūrizuoti**
  - Pasas, tapatybės kortelė, CMR važtaraštis
- **Iš dalies struktūrizuoti**
  - Sąskaita-faktūra, kasos čekis
- **Laisvos formos**
  - Sutartys, Curriculum Vitae

# Suskaitmeninti struktūrizuoti dokumentai



- El. paštu gaunami iš karto skaitmeniniu pavidalu
- Skenuojami, jei gauti popieriniu pavidalu
- Teksto atpažinimas (OCR), jei skenuoti
- Galiausiai turime dokumento tekstą +
  - Koordinates (ne visada)
  - Paveiksliukus (ne visada)

# Turinio atpažinimas



## Informacijos išgavimas

---

# Tyrimo objektas, tikslas ir uždaviniai

- Tyrimo objektas – tekstas, gautas tiesiogiai arba po OCR procedūros
- Mokslinių problemų identifikavimas ir tyrimo tikslo suformulavimas
  - Darbas ateinančiam pusmečiui
    - Tikslas stambiu planu – kaip išgauti duomenis ir informaciją iš teksto?
- Uždavinių, skirtų tyrimo tikslui pasiekti suformulavimas
- Kitų studijų metų rudens pusmetis

---

# Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

- Detalizuotas disertacijos tyrimo objektas – tekstiniai dokumentai
- Vykdoma prieinamų empirinių duomenų – dokumentų bazių, tekstynų atranka
- Suplanuotos tyrimų krypties ir disertacijos temos pristatymas pranešime 12-oje DAMSS konferencijoje 2021-12-03). Pranešimo pavadinimas - *Intelligent Data Capture in Digitized Business Documents*
- Sudaryta ir nuolat pildoma svarbiausių publikacijų preliminaria disertacijos tematika bazė. Straipsniai yra rūšiuojami, atliekama jų analitinė apžvalga

---

# Kito pusmečio darbo planas.

1. Mokslinių problemų susietų su tyrimo objektu identifikavimas ir tyrimo tikslo suformavimas
2. Mokslinių tyrimų disertacijos tema analitinės apžvalgos užbaigimas, atsižvelgiant į suformuluotą tyrimo tikslą
3. Empirinių duomenų tyrimui atrinkimas arba sugeneravimas (generavimo algoritmų konstravimas)
4. Išlaikyti privalomojo dalyko „Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika“ egzaminą (birželis)
5. Pateikti paraišką dalyvauti Rytų Europos mašininio mokymo vasaros mokykloje (Eastern European Machine Learning Summer School), liepos 6-14 d.



**Vilniaus  
universitetas**

---

# Ačiū už dėmesį

Rolandas Gricius

VU DMSTI doktorantas

[rolandas.gricius@mif.stud.vu.lt](mailto:rolandas.gricius@mif.stud.vu.lt)