



BIBLIOTHECA  
ARABICA

## ML-basierte Texterkennung arabographischer Handschriftenkataloge

### Herausforderungen und Best Practices

Konferenz „KI-Methoden im Akademienprogramm“, Hamburg 2024

Daniel Kinitz (SAW Leipzig)



Sächsische Akademie der Wissenschaften zu Leipzig





Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA

---

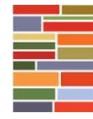


## Ablauf

- Bibliotheca Arabica: Kontext
- Herausforderungen



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



## 3 Arbeitsbereiche

Makroperspektive

u.a.  
Textgenres,  
Textpraktiken

Bsp.  
Hadith  
Commentaries,  
Randkommen-  
tare

Digitale BA

Mikroperspektive

u.a.  
MS-Vermerke,  
Bibliotheken

Bsp.  
Shirwani



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



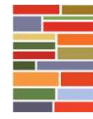
## 3 Arbeitsbereiche

Digitale BA





Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



**KHIZANA** BETA

BIBLIOTHECA ARABICA

BIBLIOTHECA ARABICA'S REFERENCE WORK ON THE ARABIC MANUSCRIPT TRADITION

KHIZANA aims to be a comprehensive bio-bibliographical reference work on agents and works related to Arabic manuscripts, focusing on the period between the 12th and the 19th centuries CE. As a reference work on Arabic literature, it integrates **sources** relevant to Bibliotheca Arabica. The most important source types are:

- data from manuscript catalogues (print & online)
- data from biographical and bibliographical works and
- manuscript notes on owners, readers, etc.

Using graph based technologies, special emphasis is laid on providing evidence (i.e., **provenance**) for every piece of information. Work on the KHIZANA will be expanded continuously in the next years. Currently, the following entries are available (including possible duplicates):

<b>PERSONS</b> 122,191	<b>WORKS</b> 120,881	<b>MANUSCRIPTS</b> 97,855	<b>MSNOTES</b> 71,287
---------------------------	-------------------------	------------------------------	--------------------------

[Sign In]

## Graph Database (Knowledge Graph)

- Data from > 100 manuscript catalogues (print & online)
- Manuscript notes on owners, readers, etc.
- ...

<https://khizana.saw-leipzig.de>

# Automatische Datenverarbeitung vs. manuelle Eingabe

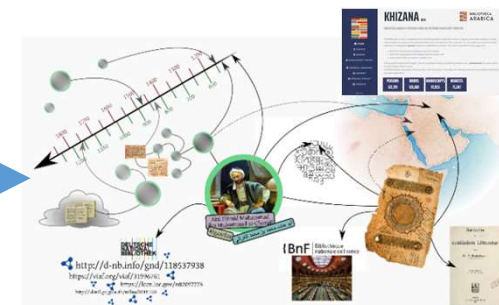
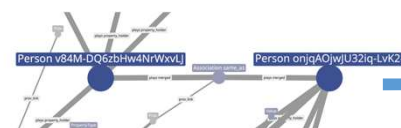
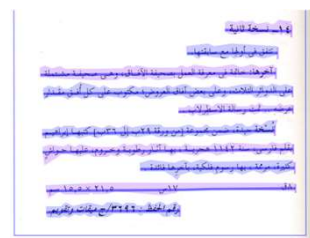
Scan (catalogue, etc.)

full text

semantic parsing

entity resolution  
(semi-automatic)

Internal / external links



Manuscript  
(Ms Notes)

full text

annotation

entity resolution (manual)



مسعود بن صالح البغدادي، و[...] المؤذن، وفضل بن سعيد بن عبيد البصر او ي:  
بن [...]، بن الشيخ بن الحسن الجيلاني، وإبراهيم بن إسماعيل بن إبراهيم الميار  
سزور الكندري، والفقير علم الدين أبي الحسن علي بن [...]. القراءة ومثبت الأمد  
عبد السيد بن [...] وذلك حادي عشر ربيع الآخر سنة اثنتين وعشرين وستمائة  
دمشق حماها الله تعالى وصلى الله على سيدنا محمد [...].

Authority Record:

- P1
- P2
- P3

# Automatische Datenverarbeitung vs. manuelle Eingabe

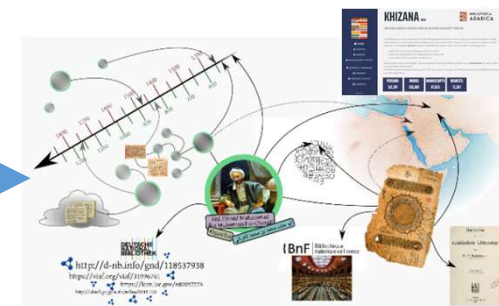
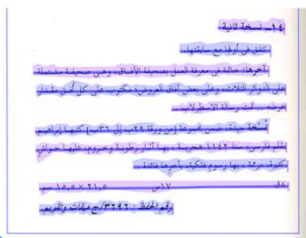
Scan (catalogue, etc.)

full text

semantic parsing

entity resolution  
(semi-automatic)

Internal / external links





# OCR: WORKFLOW



eScriptorium

kraken

image

segmentation

training/correction

full text

Line #4

كا: ملا إسماعيل، تا: ١١٨٧هـ كربلاء [مخطوطات كربلاء: ٣-٢٩٤]

كا: ملا إسماعيل، تا: 1187هـ، كربلاء [مخطوطات كربلاء: 3-294]

by Maryam (eScriptorium) on Mon Nov 27 2023 08:32:30 GMT+0100

نسخة جيدة، ضمن مجموعة (من ورقة 29 إلى 36) كتبها إبراهيم بن محمد فارس، سنة 1142 هجرية، بها آثار رطوبة وخروم، عليها حواشي كثيرة، مرصعة، بها رسوم فلكية، وآخرها فائقة .

١٧ س ١٥,٥ × ٢١,٥ سم

رقم الحفظ: ٣٦٩٦ / ج مخطوطات وتقوم.

٥٧

نسخة جيدة، ضمن مجموعة (من ورقة 29 إلى 36) كتبها إبراهيم بن محمد فارس، سنة 1142 هجرية، بها آثار رطوبة وخروم، عليها حواشي كثيرة، مرصعة، بها رسوم فلكية، وآخرها فائقة .

١٧ س ١٥,٥ × ٢١,٥ سم

رقم الحفظ: ٣٦٩٦ / ج مخطوطات وتقوم.

٥٧

نسخة جيدة، ضمن مجموعة (من ورقة 29 إلى 36) كتبها إبراهيم بن محمد فارس، سنة 1142 هجرية، بها آثار رطوبة وخروم، عليها حواشي كثيرة، مرصعة، بها رسوم فلكية، وآخرها فائقة .

١٧ س ١٥,٥ × ٢١,٥ سم

رقم الحفظ: ٣٦٩٦ / ج مخطوطات وتقوم.

٥٧

Line #4

كا: ملا إسماعيل، تا: ١١٨٧هـ كربلاء [مخطوطات كربلاء: ٣-٢٩٤]

كا: ملا إسماعيل، تا: 1187هـ، كربلاء [مخطوطات كربلاء: 3-294]

by Maryam (eScriptorium) on Mon Nov 27 2023 08:32:30 GMT+0100

- 1 آخرها: إن كان الإرتفاع شرقياً، وإلا فهو الباقي للغروب، مع زيادة
- 2 نصف التعديل في الشمال وإسقاطه في الجيوب، والله تعالى أعلم بحقيقة الحال.
- 3 نُسخة جيدة، كُتبت بقلم معتاد، في القرن الثالث عشر الهجري تقديراً،
- 4 بها آثار رطوبة وخروم، مرصعة، عليها حواشٍ، بأولها فوائد فلكية وبآخرها
- 5 نقول باللغة الفارسية .
- 6 10 ق 19 س 21×15 سم
- 7 موضوعها : فلك .
- 8 رقم الحفظ : 3825/ج مخطوطات وتقوم.
- 9 -14 نسخة ثانية
- 10 تتفق في أولها مع سابقتها.
- 11 آخرها: خاتمة في معرفة العمل بصحيفة الآفاق، وهي صحيفة مشتملة
- 12 على الدوائر الثلاث، وعلى بعض آفاق العروض؛ مكتوب على كل أفق مقدار عرضه . . تمت رسالة الاسطرلاب .
- 14 نُسخة جيدة، ضمن مجموعة (من ورقة 29 ب





Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



## Historische Ausgangslage: arabographisches OCR

- Gedrucktes Arabisch, Persisch, Urdu usw. (RTL, Ligaturen) nicht mit herkömmlichen OCR-Anwendungen gut erkennbar (hoher Aufwand, relative schlechte Ergebnisse)
  - Ab Mitte der 2010er Jahre erste frei verfügbare ML-basierte Anwendungen für OCR (kraken, Ben Kiessling), aber: zunächst ungeeignete/schlechte Transkriptionsinterfaces für RTL-Sprachen; → eScriptorium Repo seit 6 Jahren auf [gitlab.com](https://gitlab.com/eScriptorium)
- Character Accuracy: > 0.99 möglich, word accuracy: > 0.97 möglich



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA

---



# Herausforderungen & Best Practices



## Herausforderung: Kinderkrankheiten App

46	6.3			96	6 (37.20/40.30), (36.15/41.30)	
47	6.3	Geb.	37.01/31.50 tü	97		Siedl. 31.45/34.35 nhe/ he/
48	6.3			99	S. A VIII 15, A IX 7, A X 18.1, A X 19, B	
49	15.2	Siedl.	36.50/28.55	100	B IV 5, B IV 6, B IV 15, B IV 16, B VI 17	
51	artepe			98	Idüdn, Isüdn	
				102		

Rechte Spalten: Reihenfolge Zeilen vertauscht

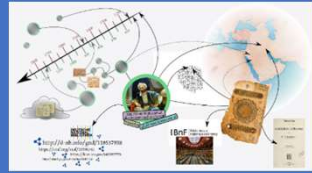
### Best Practices:

- App selbst fixen (pull request) → teuer
- auf Update warten / Arbeitspakete verschieben → umständlich
- eigenes Script → pragmatisch
- Workarounds, inklus. manuelle Arbeit SHKs



# Herausforderung: WissZeitVG

## Digitale BA



### Team

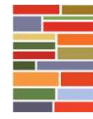
- Informatiker/RSD
- Data Engineer / Digital Humanist
- ~~IT Admin~~
- SHKs/WHKs

**Problem:** IT-Admin als wissenschaftlich-technischer Mitarbeiter → nicht mehr im WissZeitVG → „drohende“ Entfristung → Weggang

**Folge:** Verlagerung Arbeit auf mich, Informatiker, DH-Referent SAW  
+ keine enge Zusammenarbeit mehr mit kraken/eScriptorium



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



# Herausforderung: Trainingsdaten generieren → viel Arbeit

mit wenig Manpower ...

- bis vor Kurzem Fonts/Typen einzeln trainieren, von Grund auf neu
- Sep. 2022: erstes Gesamtmodell für gedrucktes Arabisch → Finetuning

The screenshot shows the Zenodo interface for a record. At the top, there is a blue navigation bar with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. Below the navigation bar, the date 'September 5, 2022' is displayed on the left, and 'Other' and 'Open Access' buttons are on the right. The main title of the record is 'Printed Arabic Base Model Trained on the OpenITI Corpus'. Below the title, the author's name 'Benjamin Kiessling' is listed. The abstract text describes the model as a text recognition model trained on the OpenITI dataset of printed Arabic-script text, available as of 2022-09-03. It mentions that the corpus includes real world Arabic (~23k lines) and synthetic data in the Tahoma (600 lines) typeface. The model was obtained by fine-tuning an Arabic-script base model on the purely Arabic-language subset of the corpus. The abstract concludes that the model is intended as a base model for fine-tuning more specific models and has not been extensively verified or optimized. The ground truth was lightly normalized to NFD but is otherwise untouched. A GitHub link is provided: [https://github.com/OpenITI/arabic\\_print\\_data.git](https://github.com/OpenITI/arabic_print_data.git). The Zenodo record ID is 7050296. To the right of the text, there is a small image of a ship and the word 'kraken'.

zenodo Search Upload Communities


September 5, 2022 Other Open Access

## Printed Arabic Base Model Trained on the OpenITI Corpus

Benjamin Kiessling

This is a text recognition model trained on the OpenITI dataset of printed Arabic-script text available at [0] in its state of 2022-09-03. It encompasses real world Arabic (~23k lines) material in a variety of typefaces augmented by synthetic data in the Tahoma (600 lines) typeface. The model has been obtained by fine-tuning the Arabic-script base model [1] on the purely Arabic-language subset of the corpus. As the model is trained on a variety of highly diverse typefaces it is mostly intended as a base model for fine-tuning more specific models from it. In line with this it has not been extensively verified or optimized. The ground truth was lightly normalized to NFD but is otherwise untouched. [0] [https://github.com/OpenITI/arabic\\_print\\_data.git](https://github.com/OpenITI/arabic_print_data.git) [1] 10.5281/zenodo.7050270

<https://zenodo.org/record/7050296>



kraken



Sächsische Akademie der Wissenschaften zu Leipzig

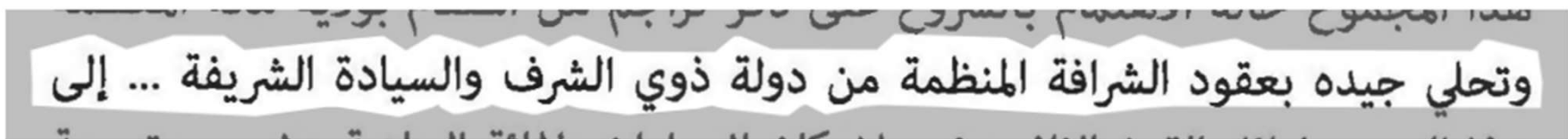


BIBLIOTHECA  
ARABICA



# Herausforderung: Korrekturlesen Hunderter Seiten mit wenig Manpower ...

→ 2 Modelle trainieren und diffs vergleichen



by admin (import) on Mon Jun 03 2024 12:01:01 GMT+0100

Toggle transcription comparison-



iraq\_indo\_digits\_vol\_1\_2\_5\_best (current)  
iraq\_mixed\_digits\_vol123\_17\_18\_20pages

وتحلى جيده بعقود الشرافة المنظمة من دولة ذوي الشرف والسيادة الشريفة ... إلى  
وتحلى جيده بعقود الشرافة المنظمة من دولة ذوي الشرف والسيادة الشريفة ... إلى



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



## Herausforderung: gemischte Schriften

→ gemischte Schriften (arabisch/persisch): Homographen und Verwechslungen pers. vs. arab. Ziffern

Bsp.: Verbundkatalog irak. Handschriften: Bd. 1-17 arab. Zeichensatz, Bd. 18 arab.+pers.

lat: 0123456789

ar: ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩

fa: ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩

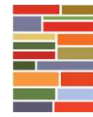
→ pers.-arab. Homographen bzw. Verwechslungsmöglichkeiten

arabisches 'Ayn (ع) → erkannt als pers. 6 (٦) → neues Modell

Herausforderung: den genauen Zeichensatz kennen (großen Textkorpus)

+ pers. Bearbeiter: arabische 1 (١ U+0661) vs. Persische 1 (١ U+06F1) → Modell trainiert verschiedene, Zeichen, die m.E. nur aus politischen Gründen versch. code points erhalten haben





# Herausforderung: Nichttrainierbarkeit der Masken

## Trainierbar:

Regionen

Baselines

## Bis dato nicht trainierbar:

Masken um Zeilen (Polygone)

معجم المخطوطات العراقية / الجزء الرابع

● تحفة المحبين شرح الأربعين النووية / شرح الحديث - عربي  
السندي المدني، محمد حيات بن إبراهيم (١١٦٣ هـ)  
المربط: الاربعون حديثاً = الاربعين النبوية = كتاب الأربعين: النووي، يحيى بن شرف (٦٣١ هـ)  
١٢٨٦٤ - بغداد: مكتبة الأوقاف العامة؛ رقم المخطوط: ١٢٤٧٨/٢  
أولها: الحمد لله - حمداً يليق به والصلاة والسلام على - صبيبه  
١٥٢٢ اسم (ف) ٢٥٣-١

وتحلي جيده بعقود الشرافة المنظمة

## Möglicher Workaround:

Statistik: seltener Endbuchstaben **و** → Postkorrektur über Liste



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



## Herausforderung: Dependencies: py bidi

→ (fehlerhafte?) Python-Implementierung für Unicode bidirectional algorithm → Zero width non-joiner (Bindehemmer) gelöscht

می‌باید mit Bindehemmer (korrekt)

مباید ohne Bindehemmer (→ inkorrektes Persisch, ggf. bedeutungsunterscheidend)

می@باید **Workaround:** durch Zeichen ersetzen, das nicht gelöscht wird

### not working for zero width non-joiner #4

🔒 Closed

Mahdizade opened this issue on Feb 28, 2016 · 1 comment



MeirKriheli closed this as completed on Jul 23 2024!

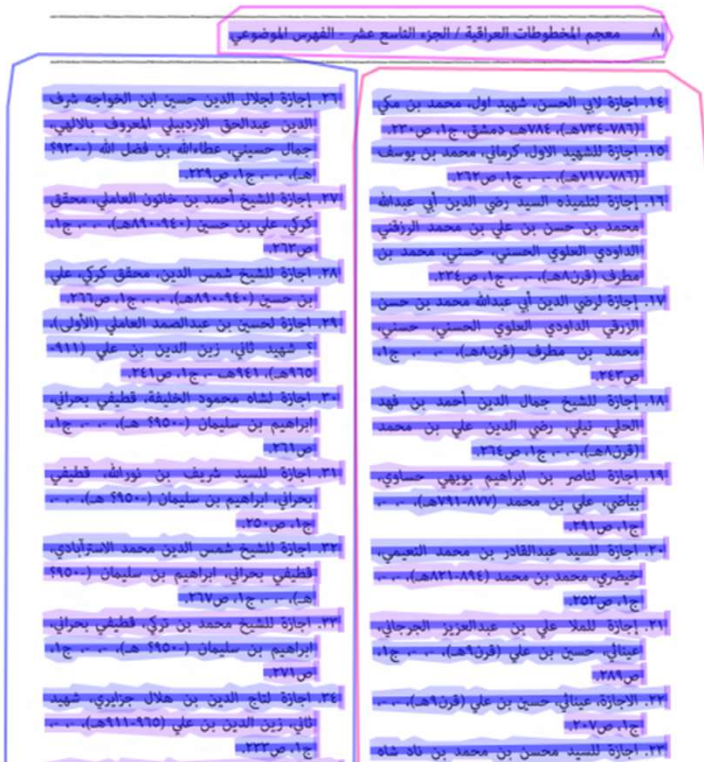
<https://github.com/MeirKriheli/python-bidi/issues/4>

→ **Problem:** Abhängigkeit von Privatinitiativen / Einzelentwicklern

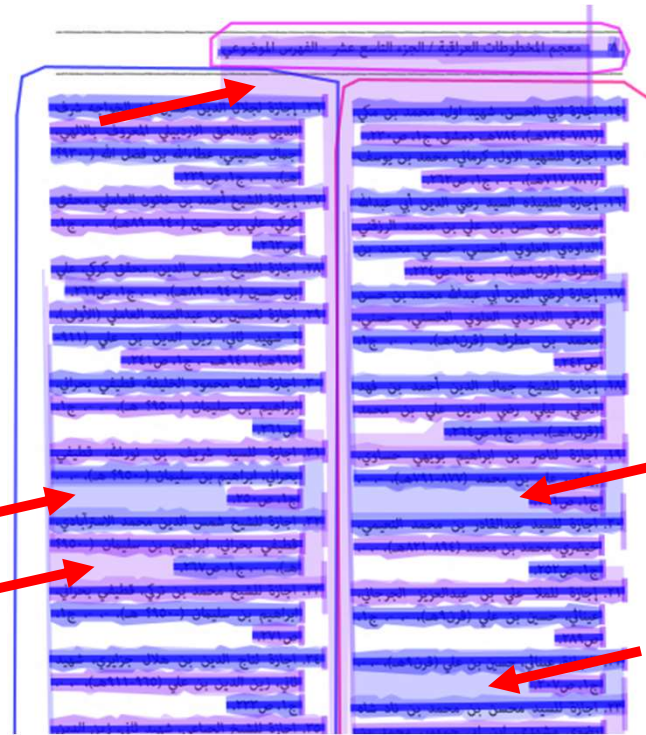


# Herausforderung: Regression

Regression bei Masken (Polygone um Zeilen), mit kranken bis dato nicht trainierbar



kraken 4  
→  
kraken 5



**Workaround:**  
kraken 4: segmentation  
kraken 5: recognition



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



## Herausforderung: manuelle Fehler + Overfitting

Bild

فأصبحنا وقد خافت يهود، ليس بها يهودي إلا وهو يخاف على نفسه.

Erkannt

فأصبحنا وقد خافت يهود، لوقعتنا بعدو الله فليس بها يهودي إلا وهو يخاف على نفسه.

Kauderwelsch  
(bestimmter Abschnitt)

?

Ground Truth

فأصبحنا وقد خافت يهود، لوقعتنا بعدو الله فليس بها يهودي إلا وهو يخاف على نفسه.

! Grund: Nachnutzung vorhandener Scan-Text-Paare (Volltexte Bücher)

**Lösungsansatz:**

Kenne deine Trainingsdaten/GT!

Kenne die Schwächen deines Modells!

Implementiere ausreichend Checks!



# Mögliche weitere Anwendungsfälle ML

- Handwritten Text Recognition (mehr Trainingsdaten)
- Normierung: identische Personen, Werke in Daten finden (s. Bild)
- Arabische Schrift  $\leftrightarrow$  latinisierte wiss. Umschrift (nicht eindeutig)
- Retrieval Augmented Generation: Sprachmodell + eigene Quellen  $\rightarrow$  Befrag den Chatbot über unsere Daten!

Entity Pairs

Max Sim: 1 Count: 555 Without Different Search History Kinitz Logout

id	Size	Sim Range	Label	Sim	Different	id	Size	Sim Range	Label
3351	1		[معين الدين مسكين بن محمد فراهي]	0.927		5266	8	0.599 - 1.000	[معين الدين محمد فرزند ترف الدين مسكين - فراهي]
338	3	0.450 - 0.923	[سيد حسين بن ابراهيم حسيني فونيني]	0.923		5276	13	0.530 - 0.950	[سيد حسين بن محمد ابراهيم حسيني فونيني]
1707	3	0.895 - 0.995	[ملا عبد الله بن محمد بههائي]	0.920		4839	1		[ملا عبد الله بن محمد بههائي]
92	28	0.365 - 1.000	[ملا ميرزا محمد بن حسن فيروكائي]	0.912		2975	1		[ميرزا محمد حائري طهرائي]
204	1		[صدر الدين محمد حسيني]	0.911		4060	1		[صدر الدين محمد حسيني]
55	6	0.716 - 0.961	[مير محمد نصير بن محمد مقصومي]	0.910		1437	1		[محمد نصير - محمد نصير بن محمد مقصوم]
837	26	0.208 - 1.000	[آقا حسين فرزند جمال الدين - خواسرائي]	0.909		4479	1		[آقا حسين فرزند جمال الدين - خواسرائي]
1223	1		[سيد محمد باقر بن زين العابدين خواسرائي اصنافائي]	0.909		3355	5	0.541 - 0.909	[سيد محمد باقر بن زين العابدين خواسرائي اصنافائي]
380	12	0.525 - 0.909	[محمد بن محمود دهقاري]	0.909		672	22	0.642 - 1.000	[محمد بن محمود دهقاري]
1388	8		[سيد حسين بن ابراهيم حسيني فونيني]	0.529 - 0.923		5276	13		[سيد حسين بن محمد ابراهيم حسيني فونيني]

Merge (M) Different (D) Merge (M) Different (D) Merge (M) Different (D) Merge (M) Different (D) Merge (M) Different (D)

id	Name	Death Date	Work Titles	Sim	id	Name	Death Date	Work Titles
1	[سيد علي بن اسماعيل حسيني فونيني]	1298	[الدرج في شرح الفرائض]	0.529	7	[سيد حسين بن محمد ابراهيم حسيني فونيني]	1208	[معارج الاكابر في شرح فرائض الاسلام ومسالك الانهار]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تحصيل الايمان في مراسم الاطهار]	0.648	4	[سيد حسين بن محمد ابراهيم فونيني]	1208	[مباهج عشاق]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تحصيل الايمان في مراسم الاطهار]	0.648	8	[سيد حسين بن محمد ابراهيم فونيني]	1300	[الانوار الزمانية]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تحصيل الايمان في مراسم الاطهار]	0.663	1	[سيد حسين بن محمد ابراهيم فونيني]	1300	[معارج الاكابر في شرح فرائض الاسلام ومسالك الانهار]
2	[مير حسينبا - سيد حسين بن ابراهيم حسيني فونيني]	1208	[مكتبة المتقنين القويدي مع الينبادي]	0.715	6	[سيد حسين فرزند محمد ابراهيم - فونيني]	1208	[معارج الاكابر في شرح فرائض الاسلام ومسالك الانهار]
2	[مير حسينبا - سيد حسين بن ابراهيم حسيني فونيني]	1208	[مكتبة المتقنين القويدي مع الينبادي]	0.773	12	[مير حسينبا - سيد حسين بن محمد ابراهيم - سينيقي فونيني]	1208	[معارج الاكابر في شرح فرائض الاسلام ومسالك الانهار]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تحصيل الايمان في مراسم الاطهار]	0.794	9	[سيد حسين بن محمد ابراهيم فونيني]	1208	[معارج الاكابر في شرح فرائض الاسلام ومسالك الانهار]
2	[مير حسينبا - سيد حسين بن ابراهيم حسيني فونيني]	1208	[مكتبة المتقنين القويدي مع الينبادي]	0.812	11	[سيد حسين بن مير ابراهيم - مير حسينبا فونيني]	1208	[مستكشف الانهار في شرح الاحكام]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تحصيل الايمان في مراسم الاطهار]	0.819	3	[سيد حسين بن محمد ابراهيم - امير ابراهيم - فونيني حسيني]	1208	[معارج الاكابر في شرح فرائض الاسلام ومسالك الانهار]

Vorschlagssystem für identische Personen (M. Reckziegel/ D. Kinitz)





Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



<https://gifer.com/de/UrpW>

Literatur:

Daniel Kinitz/Thomas Efer (2023): *Towards a Dynamic Knowledge Graph of a Non-Western Book Tradition*

In: Baillot et al. (eds.). *Digital Humanities 2023: Book of Abstracts*. Graz 2023, pp.216-217.



Sächsische Akademie der Wissenschaften zu Leipzig



BIBLIOTHECA  
ARABICA



Fragen?  
Anmerkungen?

Literatur:

Daniel Kinitz/Thomas Efer (2023): *Towards a Dynamic Knowledge Graph of a Non-Western Book Tradition*

In: Baillot et al. (eds.). Digital Humanities 2023: Book of Abstracts. Graz 2023, pp.216-217.



BIBLIOTHECA  
ARABICA

Leitung Prof. Verena Klemm



Team including research assistants of the Academy



Sächsische Akademie der  
Wissenschaften zu Leipzig

Weiterlesen