

Text Reuse and the Library Catalogue

Peter Verkinderen
Centre for Digital Humanities, AKU-ISMC



<https://docs.google.com/presentation/d/1vZH8doFmMIAaJRerSUkdQdS7c5uJJu3Fr87i94AX9oI/edit?usp=sharing>

CDH at AKU-ISMC (Est. 2023)

- Advocacy
- Methods development



(2018-2023)

- Book history
- Digital methods:
 - Text reuse
 - Citation detection



THE AGA KHAN UNIVERSITY
(International) in the United Kingdom
Institute for the Study of Muslim Civilisations

OpenITI project

(Est. 2015)

- Corpus
- OCR/HTR
- Cataloguing and metadata



AOCP

(2018-2025)

- Arabic-script OCR
- Arabic-script HTR



State of the art library catalogues...

The Digitized Collections of the Staatsbibliothek zu Berlin

Here you will find **high-quality** digitized copies of books, manuscripts and other media from the Staatsbibliothek zu Berlin. Where the originals are in the public domain, we provide them with a **public domain license**. Currently there are **216,854 works in total**.

Search within bibliographic metadata, tables of contents, or fulltexts, or **browse** a variety of content categories. In addition, there are curated collections waiting to be **explored**.

Search



Browse by:

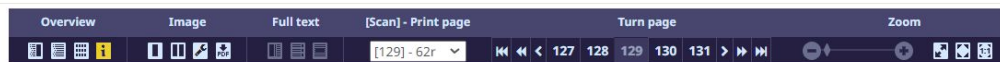
[subject \(field\)](#)[type \(material\)](#)[project](#)

Allow you to search and browse metadata...


State of the art library catalogues...

digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN1815305118&PHYSID=PHYS_0129&view=overview-info

View all metadata



Bibliographic informations

Title:	Sammelhandschrift
Year of creation:	1600
Timespan of creation:	1600 - 1699
Period of creation:	1600
Extent:	107 Blätter
Qalamos:	DE1Book_manuscript_00003312
Note:	P_Sondermat_Hss
Note:	DE1Book_manuscript_00003312
Language:	per
Language:	ara
Language:	ota
Location:	Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Berlin, Germany
Signature:	Diez A oct. 50
Category:	Außereuropäische Handschriften, Islamische Handschriften
Project :	Außereuropäische Handschriften digital
Project :	Islamische Handschriften
Digitalisator :	Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Germany
Indexing date:	Dienstag, 2. Mai 2023
Structure type:	Manuscript
Scanned Pages:	221
StaBiKat (ppn digital):	1815305118
Persistent URL:	https://resolver.staatsbibliothek-berlin.de/SBB0003381A00000000
License / Rights information:	 Public Domain Mark 1.0
METS data:	https://content.staatsbibliothek-berlin.de/dc/PPN1815305118.mets.xml
DFG viewer:	Open work in in DFG viewer

Persistent URLs



Links to digital images



State of the art library catalogues...

Full text (!?)

digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN1815305118&PHYSID=PHYS_0129&view=overview-toc

Overview Image Full text [Scan] - Print page Turn page Zoom

Compact Table of Contents

- Binding
 - Cover front
 - Paste down
 - Endsheet
- Text
 - [osmanisch-türkische Text-Stücke]
 - Stamp
 - [arabische Text-Stücke]
- Tagārib al-insān
 - Title page
 - Stamp
- Tarǧuma-i naṣā ḥ-ī ḥukamā'-i Yūnānī
 - Title page
 - Text
- Binding
 - Endsheet
 - Paste down
 - Cover back
 - Edge
 - Spine
 - Colour checker

62
71

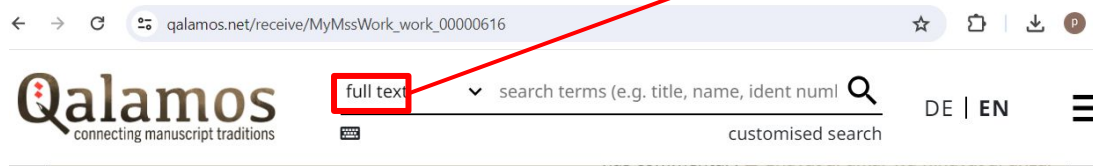
نصائح حکما حمید و سبکس و شادابی قیس افزاید که جهان را اجل جلاله و نعم نواله
با ضعاف نظرات باران دور بیک بیابان دور رود نامعدود دستوارتر
بر بجزین سستد کائنات و خلاصه موجودات اعنی محمد رسول الله صلی الله
علیه و آله و اهل بیت و یاران بعدد بزرگ و جزین و سنان جانور
کشف شد تا با بعد بداند که این رساله را از کفشار بادشاهان و حکیمان
و ائمه سلف و زبرگان و بزرگان از نصایح علما و تجار کمالان کفشار
اندک و معترفان و انصاف کرده شد تا هر که بخواند و نامالی نماید و بخرد
آورد بر همه حساب نبی و دنیاوی بیرون کرد و هم این کتاب بخارالایس
شد و الله علم سبکس نماید پادشاهی چهارچهر فایده میشود بر عدل
و دادلانم و در محاکم کبری عازم و در امور لفتت جائز و در امر بزرگ
طانم استظایس غمیر نماید چهارچهر نتوان کرد الا بچهارچهر باوشاهی
نتوان کرد که بعدل و دشمن نتوان کرد و مکوشورت و در طاصد نتوان
کرد که بتواضع و بمرا دستوان رسید مگر بصبر تقاط حکیم نماید چهارچهر پادشاهی
نگاه دارد یا کزنی دین و وزیر این و کار دشمن سیاست و نگاه داشت

Image
browsing,
download

Table of Contents

State of the art library catalogues...

Full text search (!?)



Links to manuscripts containing this text

- ▶ Works info
- ▶ Person data
- ▶ General info
- ▶ Access and usage

al-Baḥr al-zakḥkhār

Exemplars

- 🔗 [AT-ÖNB] Cod. Gl. 1: al-Baḥr al-zakḥkhār al-jāmi' li-madhāhib 'ulamā' al-amṣār
- 🔗 [AT-ÖNB] Cod. Gl. 37: Baḥr al-zakḥkhār
- 🔗 [AT-ÖNB] Cod. Gl. 38: Baḥr al-zakḥkhār
- 🔗 [AT-ÖNB] Cod. Gl. 52: Baḥr al-zakḥkhār
- 🔗 [AT-ÖNB] Cod. Gl. 62: al-Baḥr al-zakḥkhār al-jāmi' li-madhāhib 'ulamā' al-amṣār : Vol. 2
- 🔗 [AT-ÖNB] Cod. Gl. 233: Baḥr al-zakḥkhār

↓ show more

Links to parts, editions, ... of the manuscript

Link to manuscripts

- has part 🔗 [DE-SBB] Glaser 3 - 4: al-Taḥqīq fi al-ikfār wa-al-tafsīq
- has editing 🔗 [DE-SBB] Glaser 3 - 9: al-Baḥr al-zakḥkhār al-jāmi' li-madhāhib 'ulamā' al-amṣār
- has editing 🔗 [DE-SBB] Glaser 3 - 10: al-Baḥr al-zakḥkhār al-jāmi' li-madhāhib 'ulamā' al-amṣār
- has editing 🔗 [DE-SBB] Glaser 3 - 11: al-Baḥr al-zakḥkhār al-jāmi' li-madhāhib 'ulamā' al-amṣār

↓ show more

Impressive! BUT...

A lot of manual work

=> very time-consuming

=> very expensive

=> only possible for “rich” institutions

=> not easily scalable to entire manuscript collections (cherry picking)

The CDH at ISMC has been working on a number of projects to

- Make such feature-rich manuscript catalogues more feasible
- Make additional features possible:
 - Full-text search in manuscripts
 - Identifying texts in (multiple-text) manuscripts
 - Comparing and collating texts in manuscripts and printed texts

This work hinges on two main developments:

- HTR: Hand-written Text Recognition
- Text Reuse Detection: finding common passages between texts

Our plans for the future

- Integration of text reuse data in library catalogues using APIs
 - [Stanford's Digital Library of the Middle East \(DLME\)](#)
- Large-scale HTR of manuscripts
 - Single pipeline:
 - Uploading images
 - Transcription
 - Text reuse detection
 - Advantages of a corpus of manuscript transcriptions vs “critical” editions
 - Helpful for cataloguing:
 - Partial manuscripts
 - Majmū‘a (multiple-text) manuscripts
 - Suggestions for similar texts

Preparatory work: A tale of two projects

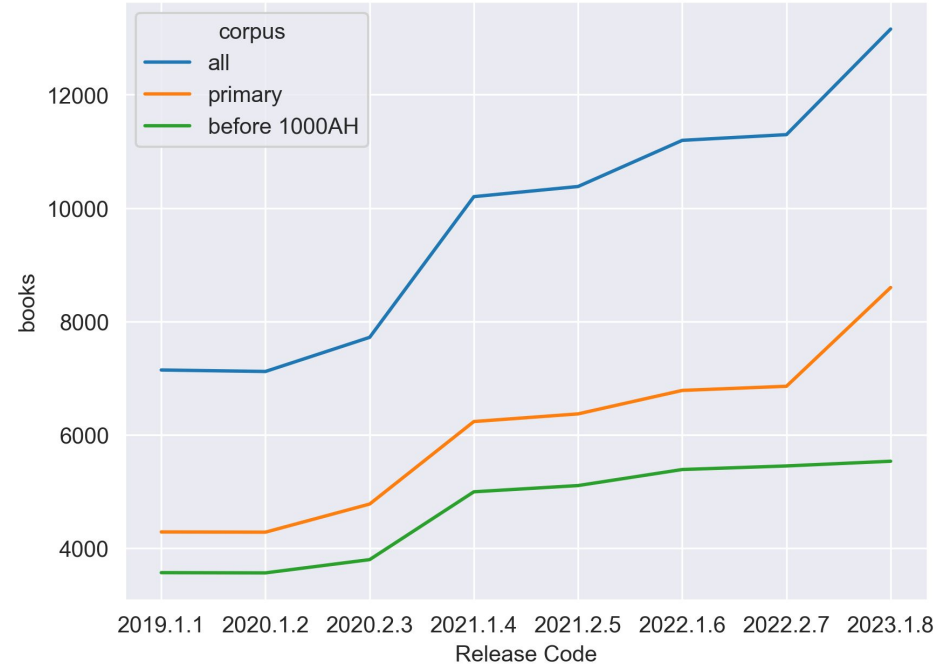
- KITAB: text reuse
- OpenITI AOCP: OCR and HTR

OpenITI Corpus

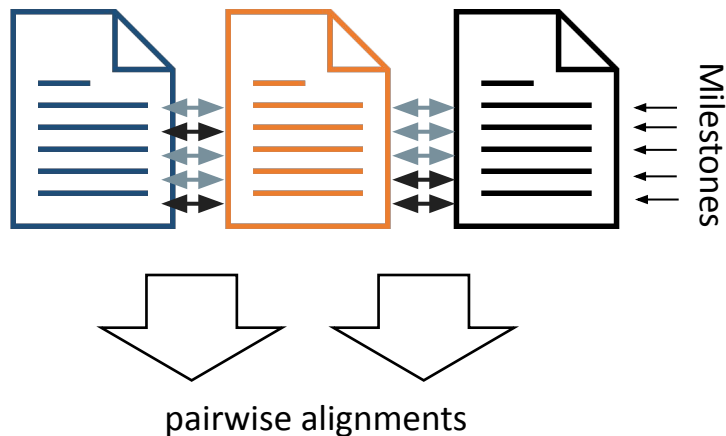
Unique Arabic Texts: 8711

With each release:

- Freeze corpus
- Publish on Zenodo
- Run passim



Text reuse detection: passim algorithm



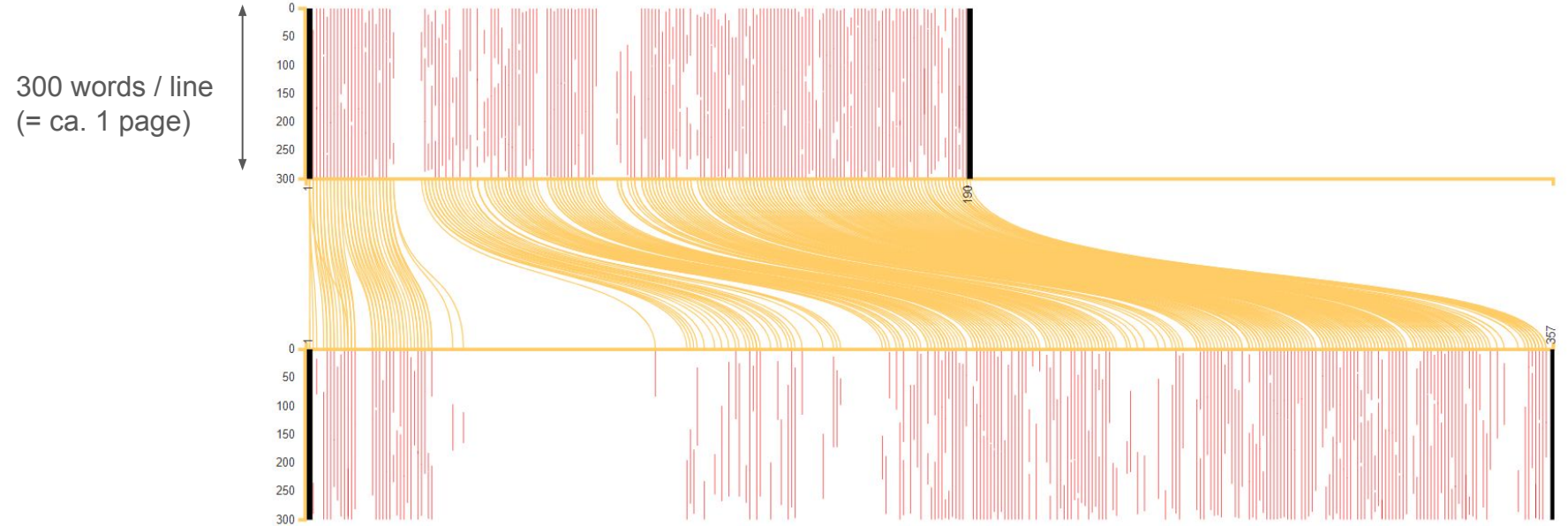
Passim : comparing over 8600 texts

1.7 Million files	3.4 Million clusters
Largest file = 32000 rows	396 clusters > 100 passages

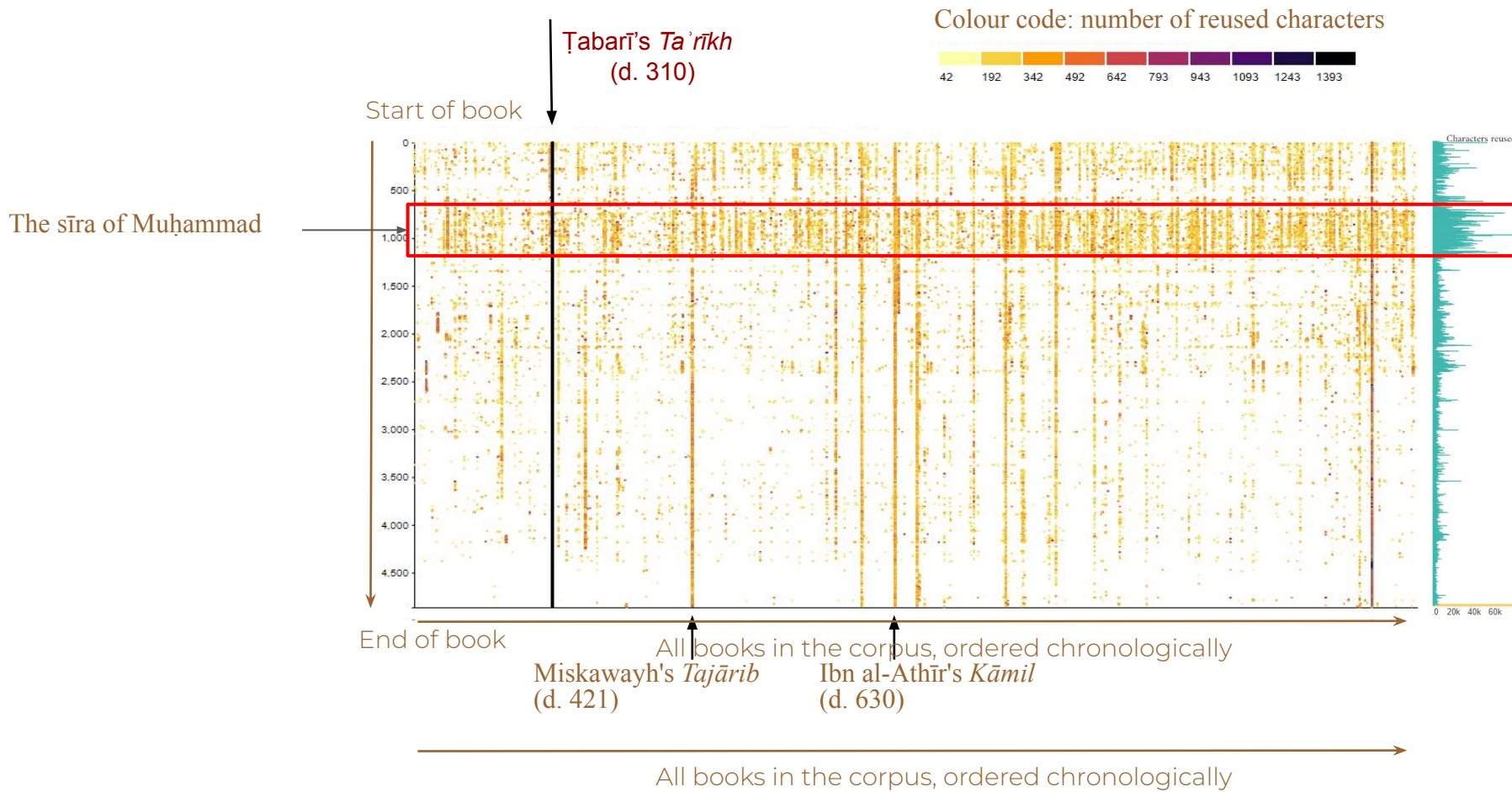
2.6 Billion Words of reuse

A5	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id1	uid1	bw1	ew1	b1	e1	len1	tok1	s1	seq1	gid1	id2	uid2	bw2	ew2	b2
2	JK001330-ara1.r-7712661770420		0	235	1	1282	1629	300	امريد جاليل مصر ومعه الانساق ا	17	-3989277145855JK000911-ara1.r-8052579093671			73	235	376
3	JK001330-ara1.r-4680357254985		72	185	370	956	1521	300	دخل في صلحهم من الروم والوثية ا	18	-3989277145855JK000911-ara1.r-6964412398730			48	87	273
4	JK001330-ara1.r-8646121275074		76	201	391	1016	1505	300	سقطع الله ظني يده ولقيها عزا حنين	65	-3989277145855JK000911-ara1.r-8494453828917			44	146	243
5	JK001330-ara1.r-8646121275074		223	294	1116	1477	1505	300	عشان بن عشان مسجد النبي صلي	65	-3989277145855JK000911-ara1.r-7289574443915			37	105	193
6	JK001330-ara1.r-7947899146155		7	289	34	1394	1448	300	هنا حلق وزحلق الباطل ومـــــــ	76	-3989277145855JK000911-ara1.r-4434787571831			6	214	31
7	JK001330-ara1.r-2768586107714		0	96	1	464	1521	300	سوقنا انه الذي اعري بعمـــــــ	77	-3989277145855JK000911-ara1.r-4434787571831			217	298	1068
8	JK001330-ara1.r-2768586107714		104	292	501	1491	1521	300	ان صاحبـــــــي اعري الناس	77	-3989277145855JK000911-ara1.r-2742603809606			2	192	14
9	JK001330-ara1.r-8363625922314		170	295	833	1483	1485	300	عن- فخره بعثت محمد وما جـ	80	-3989277145855JK000911-ara1.r-5890085293916			164	269	807
10	JK001330-ara1.r-5156852269870		0	50	1	252	1530	300	والجماعة فصار غيرو حتى وصل	85	-3989277145855JK000911-ara1.r-8136289679816			89	142	456
11	JK001330-ara1.r-6577969173989		10	213	49	1079	1525	300	عرو بن العاص وــــدخل القبطه	86	-3989277145855JK000911-ara1.r-3685436098701			0	212	1

Text reuse visualisations: pairwise



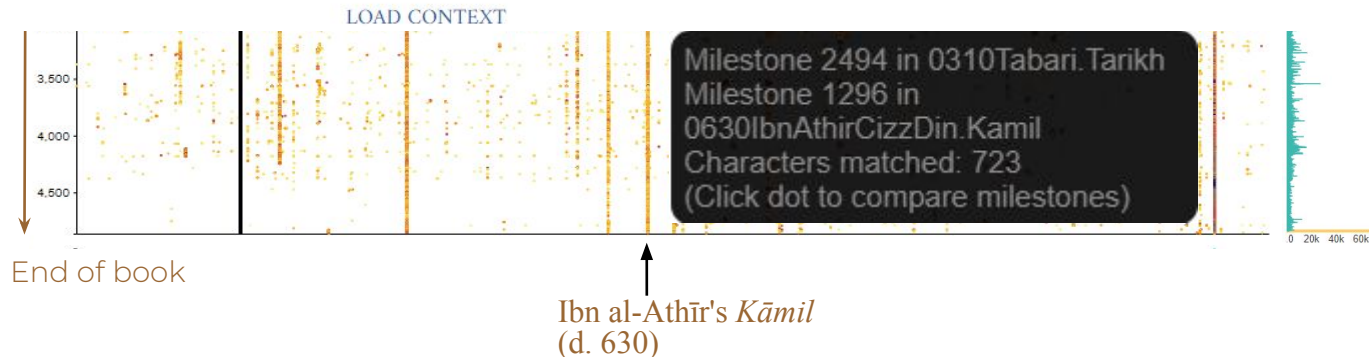
Text reuse visualisations: corpus-wide



فتحن موافك ان شاء الله للاجل الذي ضربت وفي الموطن الذي ذكرت، والسلام عليك وكتب في اسفل كتابه:
 تبصر كافي قد اتيتك معلما % علي اتلع الهادي اجش هزيم
 طويل القري نهد الشواه مقلص % ملح علي فاس الجمام ازوم
 بكل فتى لا يملا الروع تحره % محس لعض الحرب غير سووم
 اخى ثقه ينوي الاله بسعيه % ضروب بنصل السيف غير اثم
 قال ابو مخنف لوط بن يحيى، عن الحارث بن حصيره، عن عبد الله بن سعد بن نفييل، قال: كان اول ما ابتدعوا به من امرهم سنة
 احدي وستين، وهي السنة التي قتل فيها الحسين رضي الله عنه، فلم يزل القوم في جمع اله الحرب والاستعداد للقتال، ودعاء الناس في
 السر من الشيعة وغيرها الي الطلب بدم الحسين، فكان يجيبهم القوم بعد القوم، والنفر بعد النفر.
 فلم يزلوا كذلك وفي ذلك حتى مات يزيد بن معاوية يوم الخميس لاربع عشره ليله مضت من شهر ربيع الاول سنة اربع وستين،
 وكان بين قتل الحسين وهلاك يزيد بن معاوية ثلاث سنين وشهران واربعه ايام، وهلك يزيد وامير العراق عبيد الله بن زياد، وهو
 بالبصره، وخليفته بالكوفه عمرو بن حريث المخزومي، فآء الى سليمان اصحابه من الشيعة، فقالوا: قد مات هذا الطاغية، والامر الان
 ضعيف، فان شئت وثبنا علي عمرو بن حريث فانخرجناه من القصر، ثم اظهرنا الطلب بدم الحسين، وتبعنا قتله، ودعونا الناس الي
 اهل هذا البيت المستائر عليهم المدفوعين عن حقهم فقال سليمان بن صرد: لا تعجلوا، اني قد نظرت فيما ذكرتم
 فرايت ان قتله الحسين هم اشرف الكوفه وفرسان العرب وهم المطالبون بدمه ومتي علوا ما تريدون كانوا اشد الناس عليكم ونظرت
 فيما تذكرون، فرايت ان قتله الحسين هم اشرف اهل الكوفه، وفرسان العرب وهم المطالبون بدمه، ومتي علوا ما تريدون،
 وعلوا انهم المطلوبون، كانوا اشد عليكم ونظرت فيمن تبعني منكم فعلت انهم لو خرجوا لم يدركوا ثارهم. ولم يشفوا انفسهم، ولم يكوا
 في عدوهم، وكانوا لهم جزاء، ولكن بثوا دعائكم في المصر، فادعوا الي امركم هذا، شيعتكم وغير شيعتكم

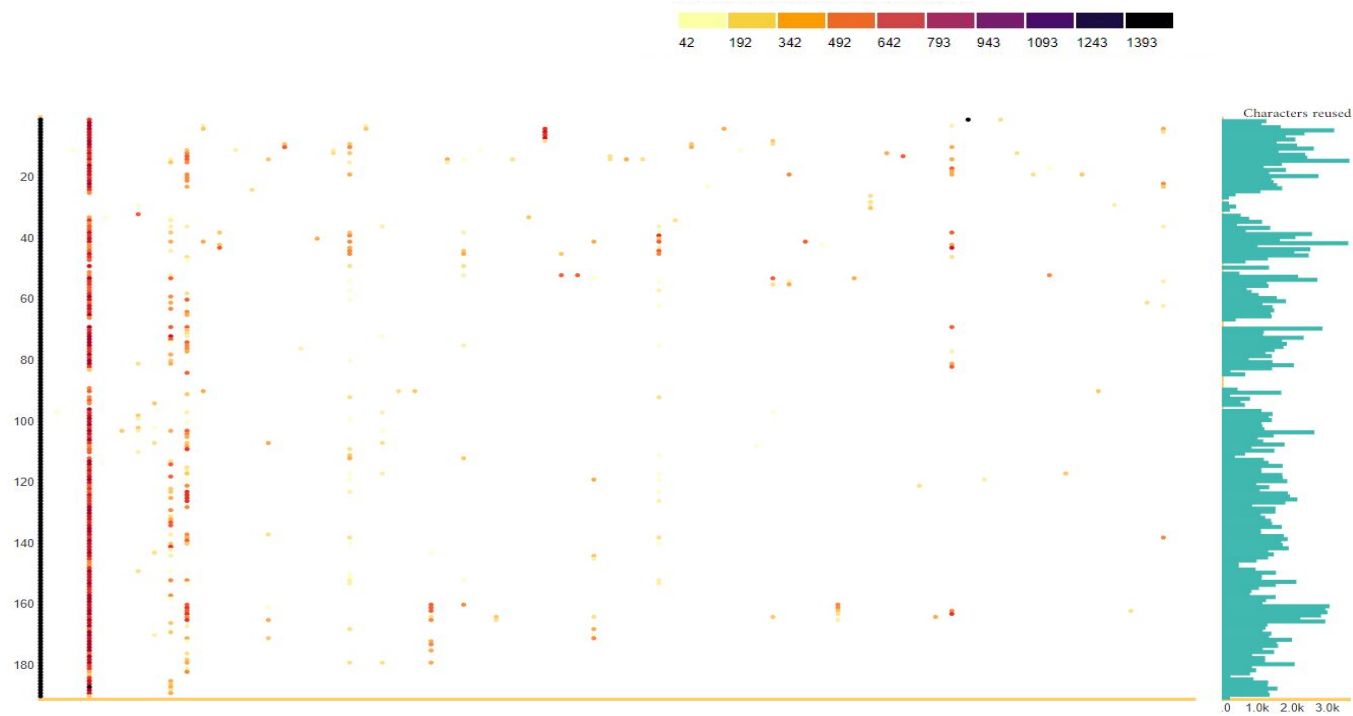
ونحن موافك ان شاء الله للاجل الذي ضربت وكتب في اسفل الكتاب
 تبصر كافي قد اتيتك معلما % علي اتلع الهادي اجش هزي
 طويل القري نهد الشواه مقلص % ملح علي فاس الجمام ازوم
 بكل فتى لا يملا الروع قلبه % محس لثار الحرب غير سووم
 اخى ثقه ينوي الاله بسعيه % ضروب بنصل السيف غير اثم
 فكان اول ما ابتدوا به امرهم بعد قتل الحسين سنة احدي وستين فما زالوا يجمع اله الحرب ودعاء الناس في السر الي الطلب بدم
 الحسين فكان يجيبهم النفر بعد النفر ولم يزلوا علي ذلك الي ان هلك يزيد بن معاوية سنة اربع وستين فلما مات يزيد جاء الي سليمان
 اصحابه فقالوا قد هلك هذا الطاغية والامر ضعيف فان شئت وثبنا علي عمرو بن حريث وكان بين قتل الحسين وهلاك يزيد بن معاوية
 ثلاث سنين وشهران واربعه ايام، وهلك يزيد وامير العراق عبيد الله بخليفه ابن زياد علي الكوفه ثم اظهرنا الطلب بدم الحسين وتبعنا
 قتله ودعونا الناس الي اهل هذا البيت المستائر عليهم المدفوعين عن حقهم فقال سليمان بن صرد لا تعجلوا اني قد نظرت فيما ذكرتم
 فرايت ان قتله الحسين هم اشرف الكوفه وفرسان العرب وهم المطالبون بدمه ومتي علوا ما تريدون كانوا اشد الناس عليكم ونظرت
 فيمن تبعني منكم فعلت انهم لو خرجوا لم يدركوا ثارهم ولم يشفوا نفوسهم وكانوا جزاء لعدوهم ولكن بثوا دعائكم في المصر وادعوا الي
 امركم هذا شيعتكم وغير شيعتكم

Interactive application:
kitab-project.org/explore



All books in the corpus, ordered chronologically

Text reuse visualisations: corpus-wide



Not all books are as well-connected...

Text reuse in the catalogue

Normalize Search Terms
 Search By Author
 Search By Book Title

Filters: Primary

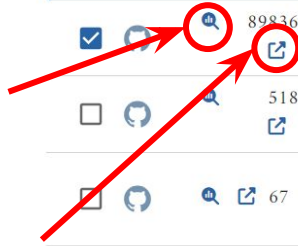
Total number of results: 6 Selected: 1 Rows per page: 10 1-6 of 6

Select a second book to visualise pairwise text reuse

More	Text Reuse	Token Count	Author Death Date	Author	Book Title	Version ID
<input checked="" type="checkbox"/>	89836 	1458897	310	al-Ṭabarī الطبري	Tārīḥ al-Ṭabarī تاريخ الطبري	Shamela0009783BK1
<input type="checkbox"/>	518 	39360	370	al-Qurṭubī القرطبي	Sila Tarikh Tabari صلة تاريخ الطبري	JK011178
<input type="checkbox"/>	67 	37091	450	Ishaq Tabari إسحاق بن يحيى الطبري الصنعاني	Tarikh San'a تاريخ صنعاء	Masaha004117
<input type="checkbox"/>	573 	58153	521	Muhammad Hamadhani الهمداني	Takmilat Tarikh Tabari تكملة تاريخ الطبري	Shamela0009783BK3
<input type="checkbox"/>	145560 	1893127	1411	Muhammad Tahir Barnaji الإمام أبو جعفر بن جرير الطبري (310 - 224 هـ)	Sahih Wa Da'if Tarikh Tabari صحيح وضعيف تاريخ الطبري	Sham19Y0145435

Corpus-wide visualization



Select pairwise visualization



















Text reuse in the catalogue

Search by author, title or attributes
Tabari Tarikh

Filters Primary ×

Total number of results: 6 Selected: 1  

Select a second book to visualise pairwise text reuse 

More	Text Reuse	Token Count	Author Death Date
<input checked="" type="checkbox"/> 	 89836 	1458897	310
<input type="checkbox"/> 	 518 	39360	370
<input type="checkbox"/> 	  67	37091	450
<input type="checkbox"/> 	 573 	58153	521
<input type="checkbox"/> 	 145560 	1893127	1411

Select pairwise visualization



Version - 2023.1.8 ×

Version URI 0310Tabari.Tarikh.Shamela0009783BK1-ara1

OpenITI Corpus Release 2023.1.8

[Visualize all text reuse](#) 

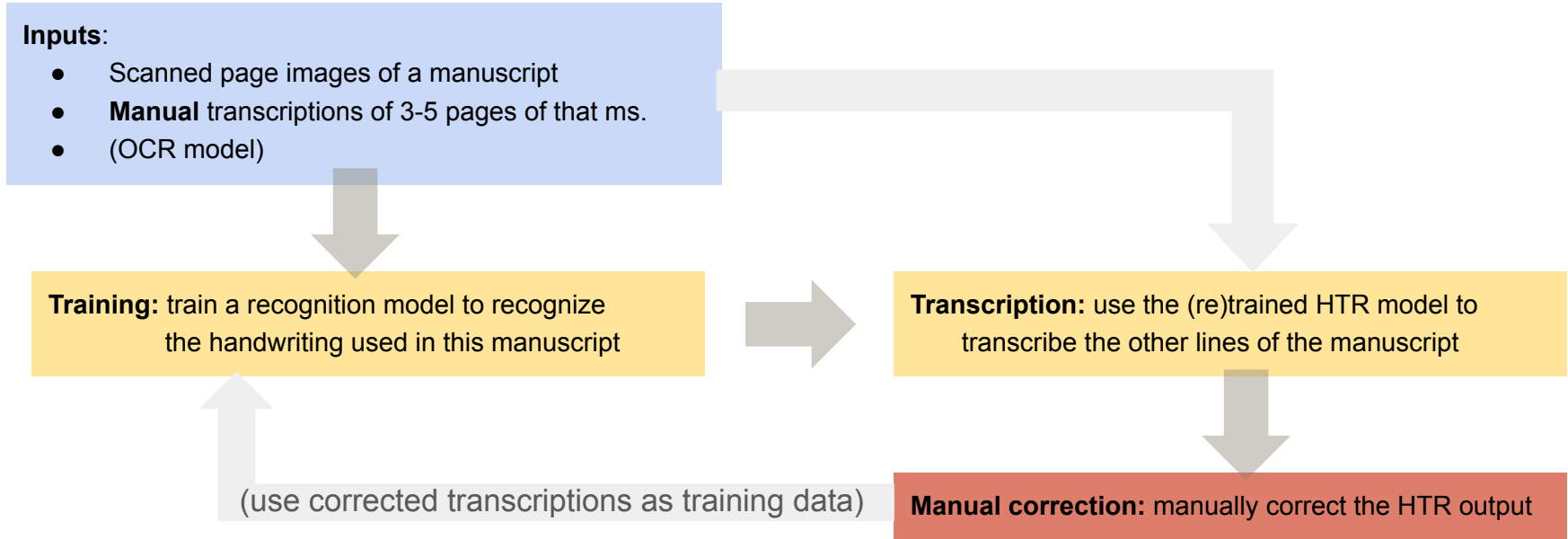
Pairwise Text Reuse Data:

Author	Title	Records #
Muhammad Tahir Barnaji	Sahih Wa Da'if Tarikh Tabari	10713  
Ibn al-Aṭir	al-Kāmil fi al-tārīḥ	4463  
Šihāb al-Dīn al-Nuwayrī	Nihāyat al-arab fi funūn al-adab	2763  
Ibn Jawzi	al-Muntaẓam fi tāriḥ al-mulūk wa-l-umam	2533  
Sibt Ibn Jawzi	Mirat Zaman	2254  
Muhsin Amin 'Amili	A'yan Shi'a	1247  
al-'Allāmat al-Maġlisi	Bihar Anwar	1058  

OCR and HTR

- OCR = Optical Character Recognition
- HTR = Handwritten Text Recognition
- AOCP (Arabic-Script OCR Catalyst Project), funded by the Mellon Foundation
- Phase 1: developing OCR models for historical Arabic-script typefaces (print)
- Phase 2: developing HTR models for Arabic-script manuscripts
- Two novel approaches:
 - Automatic Collation for Diversifying Corpora (ACDC)
 - Lacuna reconstruction

Current HTR paradigm (Transkribus, eScriptorium, ...)



Downside: a lot of manual labour, for each manuscript
=> impractical for transcribing large collections of manuscripts

AOCP Phase II - Automatic Collation (ACDC)

Step #1: User uploads manuscript images and digital edition of text

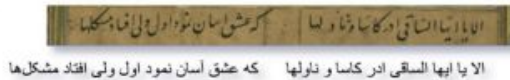
Input #1: Manuscript images of *Hāfiẓ's Divān*



Input #2: Open-access digital texts of *Hāfiẓ's Divān*

الایا ایها الساقی ادر کاسا و ناولها که عشق آسان نمود اول ولی افتاد مشکلها
به بوی ناهمای کاخر صبا زان طره بگشاید ز تاب جعد مشکینش چه خون افتاد در دلها

This line text and image are paired and become training data



This line text and image are not paired and are set aside

Step #4: ACDC tool selects top line image-transcription pair for inclusion in training data because their degree of alignment meets determined threshold and rejects bottom line due to low alignment

Step #2: ACDC tool produces a very poor (dirty) OCR transcription of the manuscript image of *Hāfiẓ's Divān*

ا ا ا ای ا ال ا اق ا بر س و نا ا ا ک ع ان مو ل و تا م ها
ب بوی نا ص ز ص ج اب تا د ا

- : used to indicate incorrectly transcribed character

ا ا ا ای ا ال ا اق ا بر س و نا ا ا ک ع ان نمود اول ولی افتاد مشکلها
ب بوی ناهمای کاخر صبا زان طره بگشاید ز تاب جعد مشکینش چه خون افتاد در دلها

Note: Successfully aligned characters are highlighted in bright green

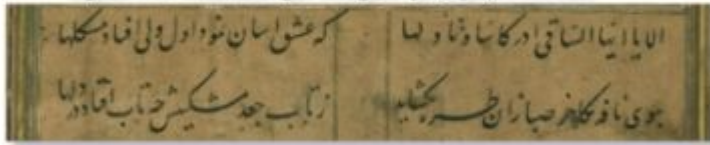
Step #3: ACDC tool aligns dirty OCR transcription of manuscript images of *Hāfiẓ's Divān* produced in step #2 with digital text (input #2) from step #1 by identifying matching patterns of characters in the two texts

AOCP Phase II - Automatic Collation (ACDC)

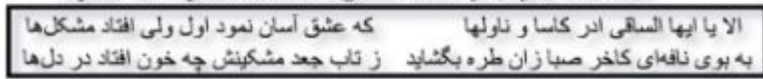
Inputs:

- Manuscript images of widely copied texts
- Digital text editions of those texts
- OCR model

Input #1: Manuscript images of Ḥāfiẓ's Divān



Input #2: Open-access digital texts of Ḥāfiẓ's Divān

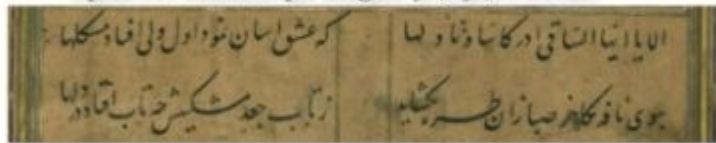


AOCP Phase II - Automatic Collation (ACDC)

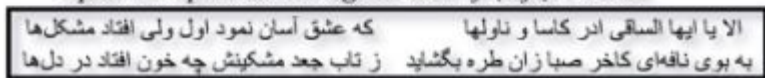
Inputs:

- Manuscript images of widely copied texts
- Digital text editions of those texts
- OCR model

Input #1: Manuscript images of Hāfīz's Divān



Input #2: Open-access digital texts of Hāfīz's Divān



Step 1: "dirty" OCR: manuscript lines transcribed with initial OCR model
=> high error rate!

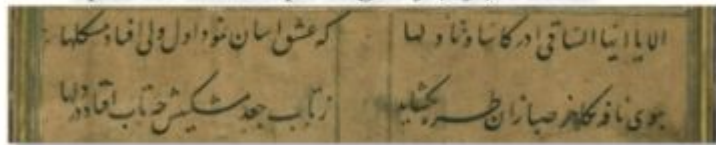
ا ا ا ای ا ال باق ا بر س و نا... کد ع ان مو ل و... ستا م... ها
ب بوی نا... هن... ز... ج... اب ستا... د...
- : used to indicate incorrectly transcribed character

AOCP Phase II - Automatic Collation (ACDC)

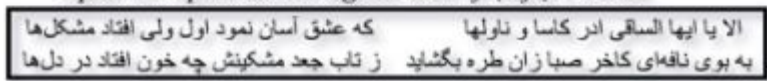
Inputs:

- Manuscript images of widely copied texts
- Digital text editions of those texts
- OCR model

Input #1: Manuscript images of Hāfiẓ's Divān



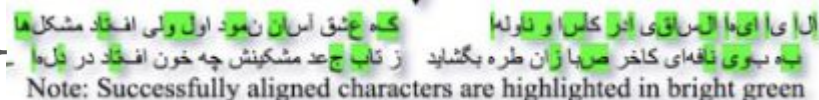
Input #2: Open-access digital texts of Hāfiẓ's Divān



Step 1: "dirty" OCR: manuscript lines transcribed with initial OCR model
=> high error rate!



Step 2: alignment: use a text reuse algorithm to align OCR output to digital text editions



AOCP Phase II - Automatic Collation (ACDC)

Inputs:

- Manuscript images of widely copied texts
- Digital text editions of those texts
- OCR model



Step 1: “dirty” OCR: manuscript lines transcribed with initial OCR model
=> high error rate!



Step 2: alignment: use a text reuse algorithm to align
OCR output to digital text editions



Step 3: retrain OCR model: select the n lines that were
aligned best to texts editions and use the aligned text from the
editions (!) to retrain the OCR model

AOCP Phase II - Automatic Collation (ACDC)

Inputs:

- Manuscript images of widely copied texts
- Digital text editions of those texts
- OCR model

Step 1: “dirty” OCR: manuscript lines transcribed with initial OCR model
=> high error rate!

Step 2: alignment: use a text reuse algorithm to align
OCR output to digital text editions

Step 3: retrain OCR model: select the lines that were aligned
for more than n % to texts editions and use the aligned text
from the editions (!) to retrain the OCR model

Step 4: re-OCR: use the retrained model to OCR the
other manuscript lines

Repeat a few times

Advantages:

- no manual training data generation
- model reusable across manuscripts

Results:

- Increased transcription accuracy:
average 80% instead of 60% !

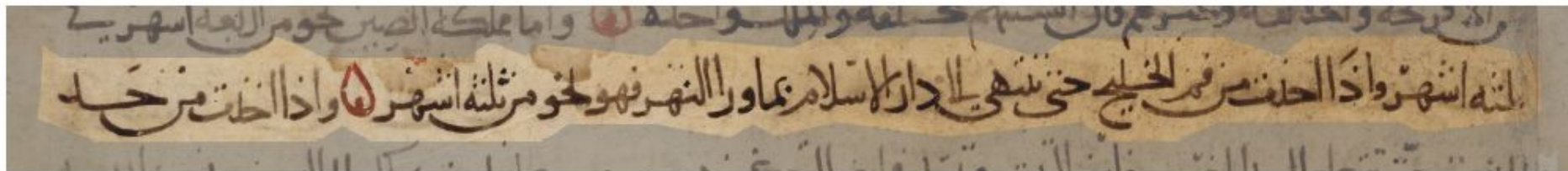
A line of text from a manuscript in an unseen hand, transcribed by the model generated by David Smith using the ACDC method (without any manual training data):



الذي يأخذ من هذا البحر المحيط من أرض المغرب وبارض الأندلس فقد قسمت هذا الأرض قسمين وخط هذه القسمة

الذي يأخذ من هذا البحر المحيط من أرض المغرب وبارض الأندلس فقد قسمت هذا الأرض قسمين وخط هذه القسمة

A line of text from a manuscript in an unseen hand, transcribed by the model generated by David Smith using the ACDC method (without any manual training data):



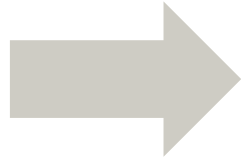
ثلاثة أشهر و إذا أختت مر فم الخليج حتى ينتهي لأدار الا سلام بماوراء النهر فهو نحو مرثلاثة أشهرها وإذا أختت من حد



AOCP Phase II - Lacuna Reconstruction

Stage 1 : Mask out parts of the text - train to recognise masked parts (train on 1000s of folios!)

Stage 2 : Finetune on a small number of transcriptions (if needed)

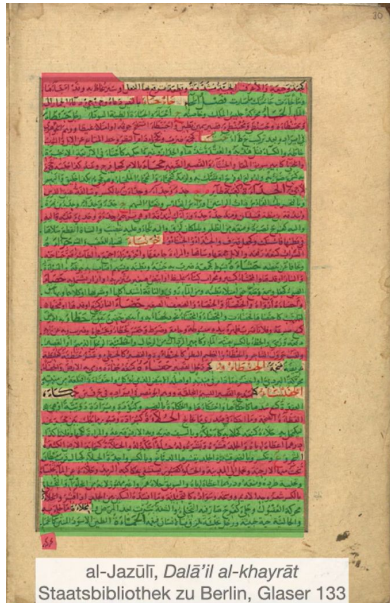


```

VPOS="860"
WIDTH="2075"
HEIGHT="762">
<Shape>Polygon POINTS="2498 986 2448 901 2423 881 2342 860
2337 860 2332 860 2327 860 2297 886 2261 860 2246 866 2211 866 2186 891
2035 871 1994 876 1944 876 1894 876 1843 876 1793 876 1743 881 1692 881
1642 881 1591 881 1541 881 1526 881 1460 901 1415 886 1390 886 1340 886
1324 886 1284 906 1284 911 1128 911 1027 911 987 891 987 891 937 896 921
896 901 911 896 911 750 896 735 896 685 901 634 901 584 901 549 901 433
916 438 1032 438 1082 715 1122 730 1122 725 1122 730 1122 785 1087 891
1117 896 1117 901 1117 1108 1087 1299 1112 1304 1112 1531 1072 1607 1107
1612 1107 1617 1107 1622 1107 1627 1087 1632 1082 1924 1092 1929 1092 1934 1092
1939 1092 1974 1062 2025 1097 2030 1097 2035 1097 2040 1097 2045 1097
2050 1097 2050 1092 2075 1072 2191 1097 2196 1097 2201 1097 2302 1052
2508 1047 2508 996 2498 976 2498 986"/></Shape>
<String CONTENT="حدث في الارض فيما هذا السنة ووضعا فيه هذه">
HPOS="433"
VPOS="860"
WIDTH="2075"
HEIGHT="262"></String>
</TextLine>
<TextLine ID="eSc_line_51239498">
BASELINE="446 1212 2526 1175"
HPOS="438"
VPOS="1057"
WIDTH="2085"
HEIGHT="226">
<Shape>Polygon POINTS="443 1208 443 1260 1324 1263 1329 1293 1283
1586 1253 1627 1248 1642 1248 1727 1273 1732 1273 1738 1273 1743 1273
1748 1273 1783 1248 1788 1248 1808 1248 2141 1268 2146 1268 2518 1233
2523 1173 2513 1087 1924 1057 1919 1057 1914 1057 1838 1092 1607 1067
1602 1067 1596 1067 1551 1067 1521 1077 1516 1077 1511 1077 1517
1249 1102 1168 1072 1163 1072 1158 1072 1153 1072 1148 1072 1148 1077
1118 1102 992 1082 987 1082 982 1082 977 1082 947 1107 735 1077 738 1077
438 1117 443 1208"/></Shape>
<String CONTENT="الفلان فرغت عصر وانصبت حتى عد كل ارضه بالقي">
HPOS="438"
VPOS="1057"
WIDTH="2085"
HEIGHT="226"></String>
</TextLine>
    
```


AOCP Phase II - state of the HTR art

- HTR accuracy is already much higher than anticipated
- Main bottleneck (unexpectedly!): page segmentation: extracting line images from page images is tricky (slanted lines, vowel markers, diacritics, ...)



al-Jazūlī, *Dalā'il al-khayrāt*
Staatsbibliothek zu Berlin, Glaser 133

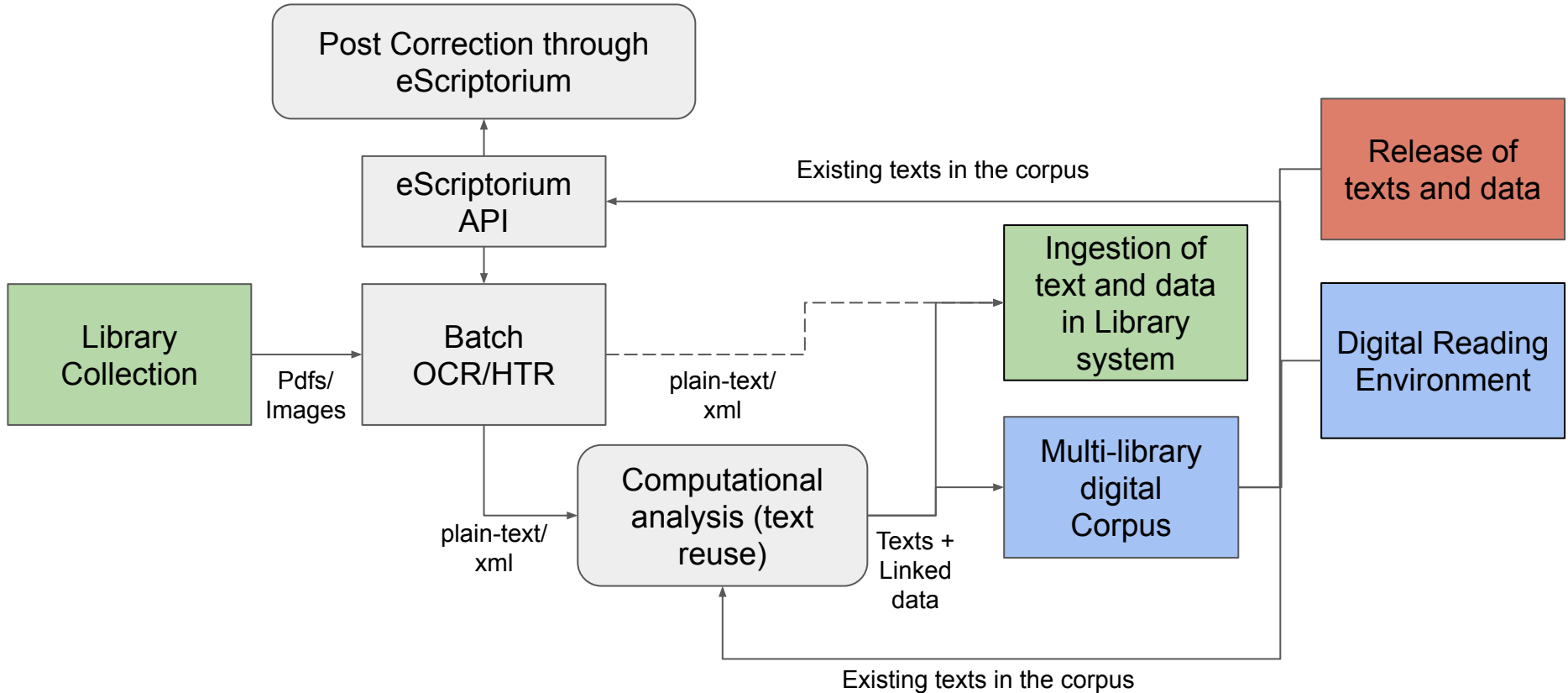


Sa'dī, *Gulistān*
Library of Congress PK6450 .G2 1593

Our plans for the future

- Integration of text reuse data in library catalogues using APIs
 - [Stanford's Digital Library of the Middle East \(DLME\)](#)
- Large-scale HTR of manuscripts
 - Single pipeline:
 - Uploading images
 - Transcription
 - Text reuse detection
 - Advantages of a corpus of manuscript transcriptions vs “critical” editions
 - Helpful for cataloguing:
 - Partial manuscripts
 - Majmū‘a (multiple-text) manuscripts
 - Suggestions for similar texts

Large-scale automated HTR and libraries



Proof of Concept grant

- Talk to libraries and manuscript holdings to
- Implement the pipeline
- Test the pipeline on a number of collections