

# Handwritten Text Recognition

$u^b$

b  
**UNIVERSITÄT  
BERN**

Tobias Hodel (Walter Benjamin Kolleg, Digital Humanities)

Bern, 04.05.2023

# Goals

We want to...

- understand the basics of Text Recognition and be able to produce specialized models on our own
- develop scenarios for the use of Text Recognition for your research
- get to know export functionalities and ways to publicize transcriptions
- get to know challenges for project based on machine learning

# Before we start

Please:

- > Register at <https://readcoop.eu/>
- > Download the latest version of Transkribus:  
<https://readcoop.eu/transkribus/download/>

There is a lite (web-version) available: I recommend to learn the (full) downloadable software (if you want to use Transkribus for research).

Mac User: Make sure to copy Transkribus to your program folder.

Java-Problems: [https://download.oracle.com/java/17/latest/jdk-17\\_macos-x64\\_bin.dmg](https://download.oracle.com/java/17/latest/jdk-17_macos-x64_bin.dmg)

# A very short overview: From Optical Character Recognition to (Automatic/Handwritten) Text Recognition

1990ies: Commercial Optical Character Recognition

2000thes: Recognition of «Fraktur» with OCR (EU Projekt IMPACT, ABBYY XIX)

2010ns: Simple Language Model / Intorduction of neural networks

2020ies: Dominance of *deep learning*  
(high-dimensional language model)

Starting 2021: Transformer-basierte Modelle für Texterkennung (trOCR)

# Projekt READ

## Recognition and Enrichment of Archival Documents

Making archival (esp. handwritten) documents more accessible

Research infrastructure – Transkribus

Funded until mid-2019 by the European Union (H2020)

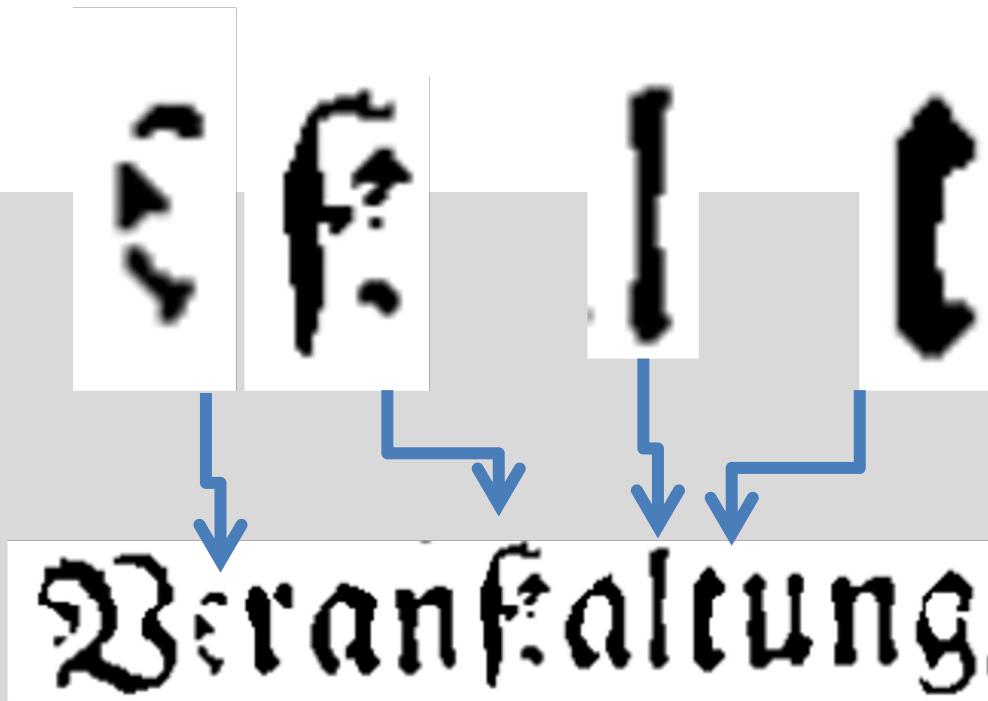
15 European partners

# Research perspectives of READ

- Recognition of layout and text structures
- Recognition of handwriting (Handwritten Text Recognition)
- Text recognition with dictionaries
- Writer identification --> proof-of-concept
- Best-practices for recognition of large amounts of documents
- Digital Humanities in archives and scholarly practices

**READ**

# Automated Text Recognition?



# Automated Text Recognition

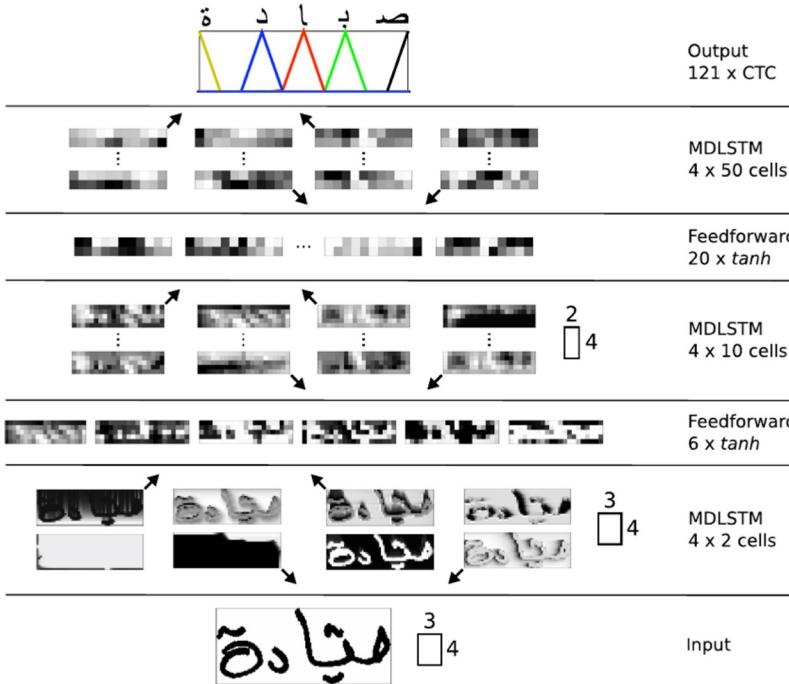
- Machine learning using neural networks
- Processes writing by line, rather than by character
- Needs to be trained by being shown document images and transcripts
- More training data → more accurate recognition
- Create a model to transcribe and search a collection of documents

# Recognising printed text

- . Neural networks can also process printed text – with less training data!
- . Transcribe documents or use OCR engine in Transkribus
- . Use these transcripts to train a model
- . Results with <1% CER are possible



## Von der Eingabe zur Netzausgabe



(Bild aus GRAVES, SCHMIDHUBER: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks)



# Carolingian Minuscules

5'000 words, 1'000 lines:

Validation pages: 7% CER

24

illecm: que p[ro]fec[ta]cc fo[rum] mitat  
timorem per quic uiuersc que  
prius non sine formidine obser-  
vabec absq[ue] ullo labore uel ut  
nec tunc eliter & consu&udine  
incipi & custodire. Non icem timo-  
re gehenn[us], sed amore xp[ist]i: & con-  
su&udine ipse bone & delecta-  
tione uirtutum: quae dñs icem  
in operchum suum mun dum  
auitatis & pec[un]ia[rum] sc̄o dignebit[ur]  
demonstrare; **De officiis diuinis** **III**  
**H**iemit tempore **IN HOC TIBUS**,  
idesat: alcalen dis nob[is] bem brit[us]. in festo marie panis  
usque in pascha. iux ta considera-  
tionem rectionis o[ct]ava horee noc-  
tis surgen dum est: ut mo di ce-  
am plius de media nocte pecus &

# Early Modern Letters

48'000 words, 6'450 lines:

**Validation pages: 2,5% CER**

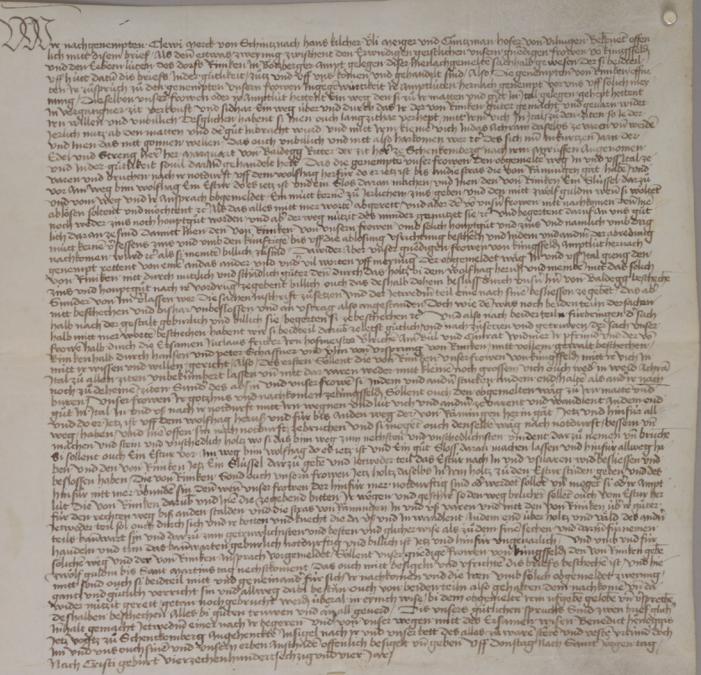
allor lind by lar St. venthar Dreyfus 27

mit was foriden und was gretig an habe sol  
gleichsich van dreyt enz. dies geschenk an  
vader gott seg lob und dank da stadt unter  
der sonden salt der sol gro kommen zet  
der wyls doy gow in den vryd wahr  
nun wond gow in dant, der geygafft  
zu vryl ghet will ich acht uoud zelten  
fall vnde den dor ouys si sol kommen  
wys min inticht dorvan aben und reicht  
hiedam doy sol nicht anys balt favord  
si ouys my dor, wo den ic min leid den  
mechewob regens will zelten sin vnde  
israe sogen so den day, si ouys acht seide  
aldeinde regenten deuon, si geban enghen en  
my dor want nglieh laugen gemit  
da feuerst und wird alldeinen in sic gelegten  
es ghan vor vryt gemit si ghan ic fort  
dag dor my nicht, dorwan my dor  
maghe, isam vordern speltum, in miche  
ouys endem dor von vher da vnde auf wa  
gekomen vnder agen, wie sollan dor

A.

# Charter 15th century

77'000 words, 3'500 lines:  
**Validation pages: 4,8% CER**



# Council Minutes 19th Century

148'000 words, 31'000 lines:

**2,5% CER**

14. Januar 1851.

Oktum Samstags den 4. Januar 1851.  
In Gegenwart des gesammten Regierungsrathes.

so wird diejenige Aktion des Körpers, die für  
mich interessant ist, dem Prinzip der Aufrech-  
tstellung und Bewegung zu Entstehung und  
Verlusten des Gleichgewichts zu unterwerfen.

Opfördert die Organisat. für ca. 28.000 Katholiken. Von demselben Organisat.  
betreut werden die Katholiken der Stadt und des Landes. Das Organisat. hat seinen Sitz in der  
Stadt Breslau.

Dorffürstl. von Stobn. 400 an das Bezirksgericht W<sup>o</sup>  
 wird der Vierblatt des Einwohner 23. Febr. 18.

gäng überwipf.

Allesamt gelblichgrün,  
Habitus der Art Agave  
mit dichten, blattartigen  
Blättern, die an den  
Blütenstand heranreichen.  
Blütenstand mit langer  
Röhre, oben zu einer  
Kugel ausgebildet,  
die aus zahlreichen  
kleinen Blüten besteht.

# Generalizing of Writing Types Current/19th century

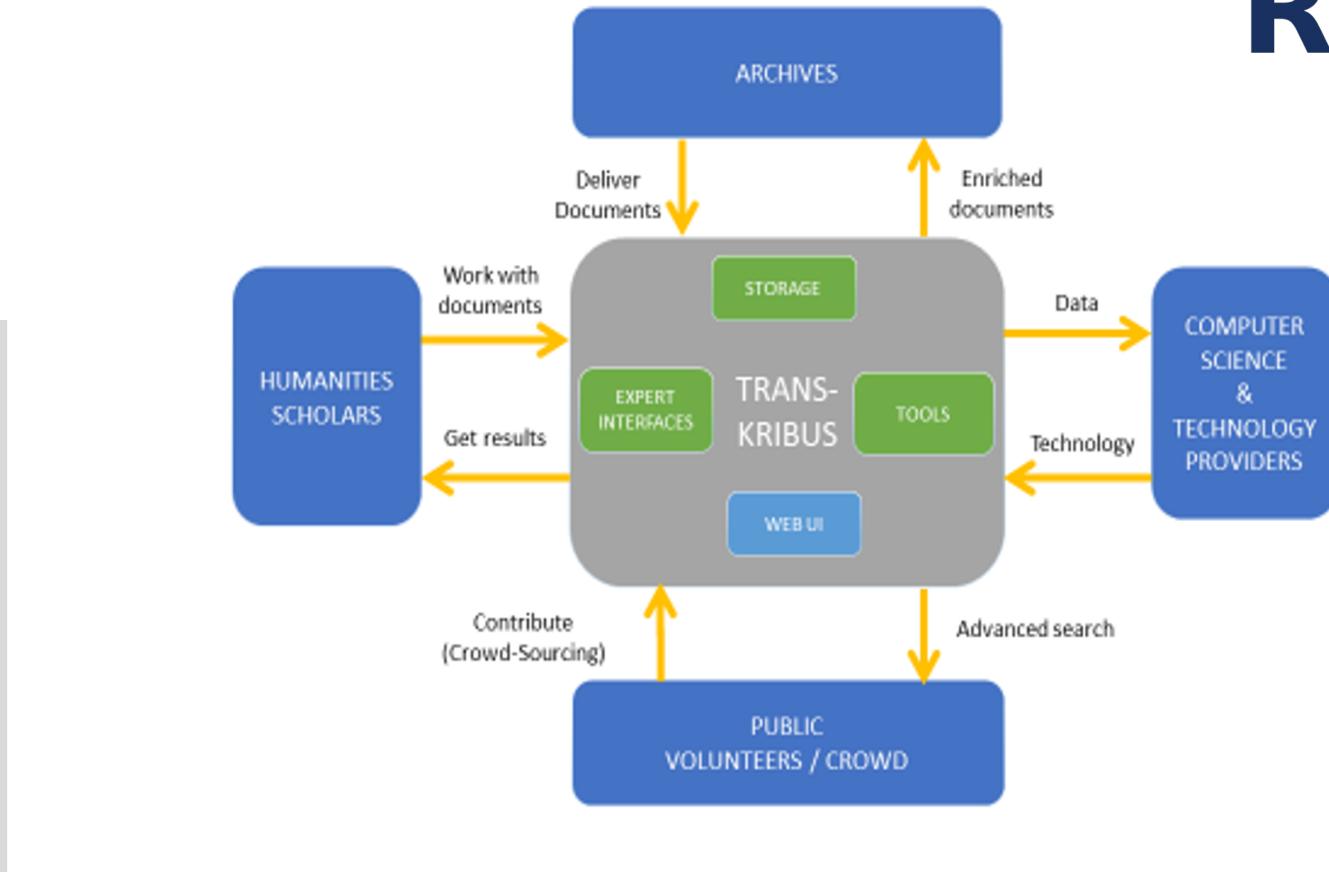
HTR MODEL	HTR ENGINE	CER MEAN %	CER MEDIAN %	CER UPPER BOUND (WORST)
German Kurrent M2	HTR+	3.43	2.76	9.13
	PyLaia	18.77	13.30	51.05
Transkribus German Kurrent	HTR+	5.90	4.85	10.20
RRB	HTR+	9.15	8.13	16.28

1. Januar 1851.

Actum Suntays den 4. Januar 1851.  
In Gegenwart des gesammten Regierungsrathes.

Die bei Hoffnung auf Vorschriften von 30.000 aufgelegte Summe ist im Konsolidationsfond der fünfzehn Kantone eingezogen, bis gegen Ende des Jahres 1851 zu vertheilen. Die Konsolidationsfond wird aus dem Konsolidationsfond der Kantone, welche am 31. März 1850 aufgelegt sind, und aus dem Konsolidationsfond der Kantone, welche am 1. Januar 1851 aufgelegt sind, zusammengestellt, und die

Zur Hand zu nehmen, dass, um Abstimmungen in Bezug auf das Gesetz, welche die Konsolidation des Konsolidationsfonds der Kantone, welche am 1. Januar 1851 aufgelegt sind, auf die von den Kantinen



*u*<sup>b</sup>

Look at Transkribus

---

<sup>b</sup>  
UNIVERSITÄT  
BERN

**READ**

# READ COOP

## Business model

Cooperative

More than 60 members

(Institutes & private members)

Buy-in:

250€ / 1'000€

500 Credits included

Costs of Credits: Varying depending on amount.

<https://readcoop.eu/transkribus/credits/>

Ca. 0.11-0.24 € / page (text recognition)

# Visual Feature Recognition

page-number

marginalia paragraph

320.

Daudor fuhkend und den Rita  
Loutinum das grose Gabium u. Zalpa  
in der fl. Und zu Daudor Basel un  
königlich. Und zu H. singlangtan da  
= gebra, und gema' d und dor  
= galangtan, Gabium das nu  
= gern d'riovalor domisior  
H. Daudor H., soll das in da

page-number

marginalia paragraph

321 preamble

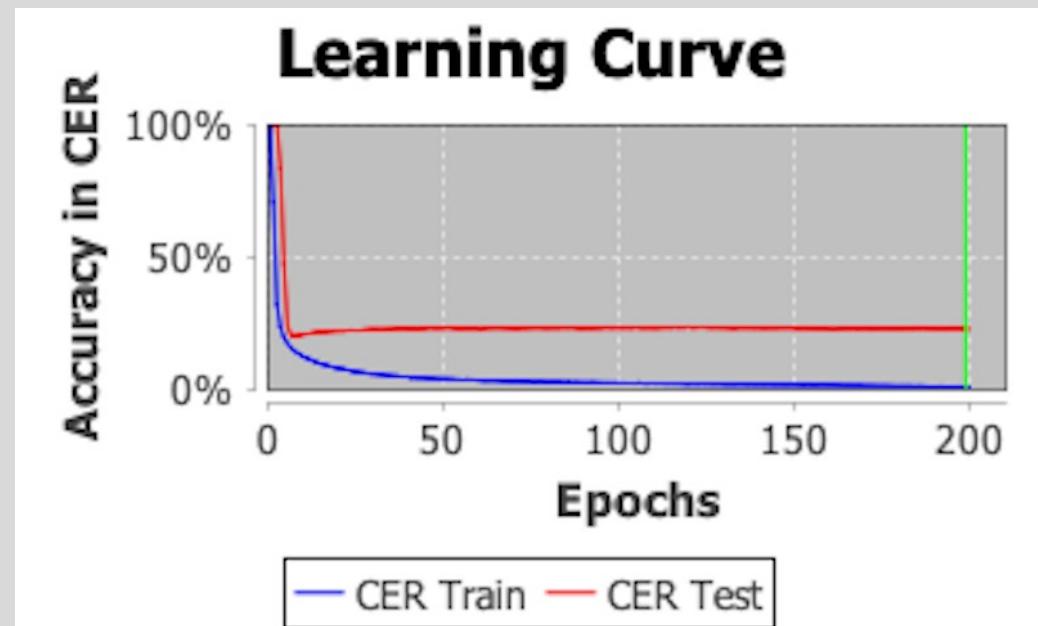
Von für mit Rauf Enföning wo hör ich  
Girolf und H. Narfingar mit Girolf  
lex s. offn. Willam von Hessen aufgenommen  
Karföre, warhier unten anhören  
Herr P. Ich will nun siehe  
meistens zu Blattföllen be  
gungen sind, - haben Magie  
an die darüber liegenden hundert

# Visual Feature Recognition

	<p>1461. Februar 9. (Mittag vor der pfaffen vasmacht.)</p>	<p>Werner Hofer, Gerichtsschreiber der Stadt Zürich, erklärt, da Oldelheit Völlinger von dem Predigerkloster Zürich ein Häuschen an dem kleinen Predigergässchen an dem Brach zu seinem länglichen Niesbruch gekauft habe, mit der Bestimmung, dass aus ihrem Nachlass 20 Pfund Penninge an das Kloster bezahlt werden sollen, so habe er, nachdem ihm dieselbe all ihr Gut versprochen habe, dem Kloster Gewähr geleistet, dass, wenn der Nachlass für Bezahlung der 20 Pfund nicht hinreichen sollte, es das Mangolende ergangen werde.</p> <p>Hofer siegt.</p>
call-number	<p>St. A Z. Obmannamt str. 286. Original, Pergament. Siegel hängt.</p>	

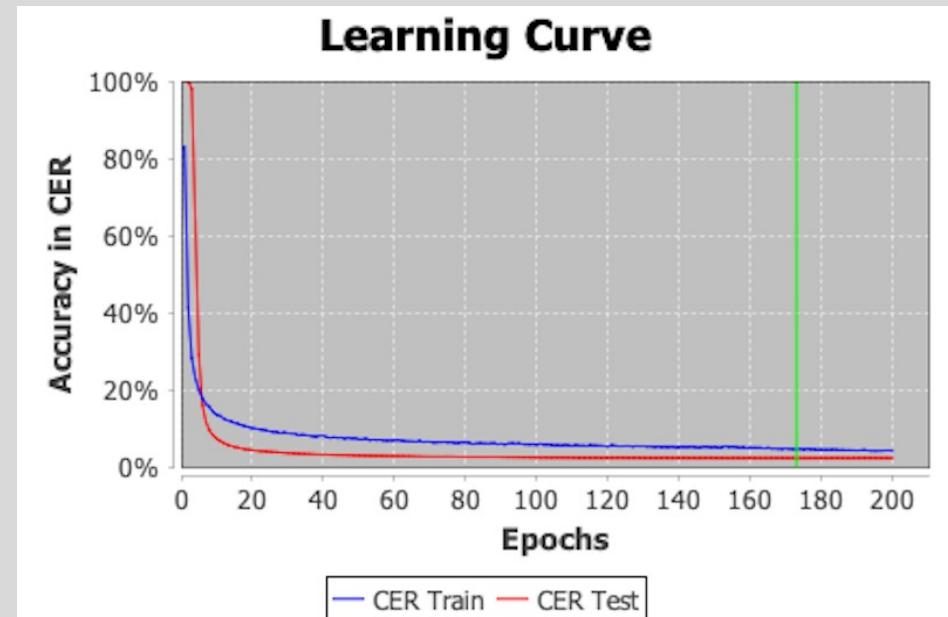
# Interpretation of Graphs

Diverging curves: overfitting



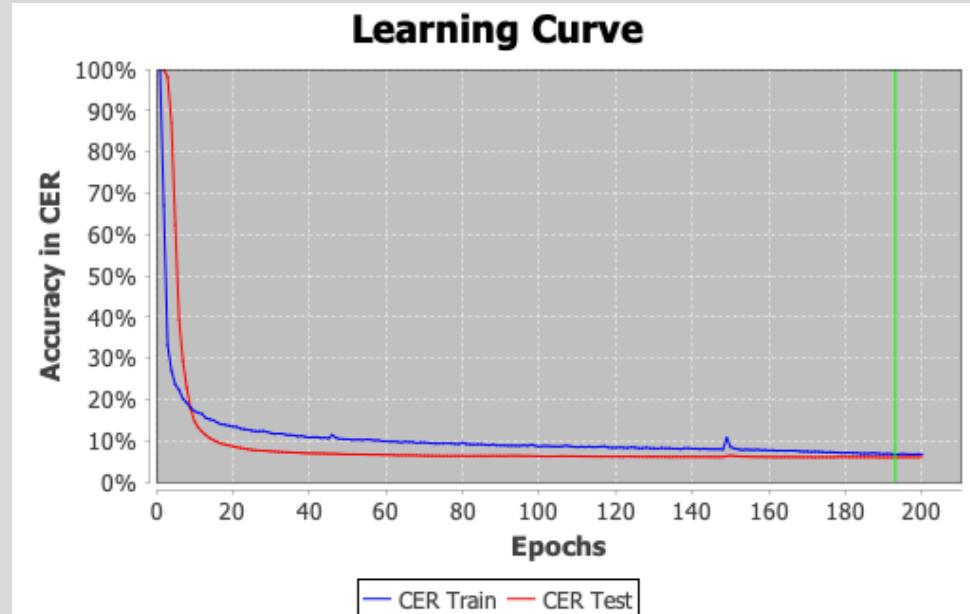
# Interpretation of Curves

Keep on training...



# Interpretation von Trainingskurven

## Austrainiertes/optimales Netz



## How-To's & Links

Read&Search (Hanserezesse)

<https://transkribus.eu/r/rezesse-niederdeutscher-staedtetage/#/>

Tutorials:

<https://readcoop.eu/transkribus/resources/how-to-guides/>