Zeitschrift des Max-Planck-Instituts für europäische Rechtsgeschichte
Journal of the Max Planck Institute for European Legal History

Rechts Rg geschichte

Rechtsgeschichte
Legal History

www.rg.mpg.de

http://www.rg-rechtsgeschichte.de/rg27
Zitiervorschlag: Rechtsgeschichte – Legal History Rg 27 (2019)
http://dx.doi.org/10.12946/rg27/244-259

Rg 27 2019 244 – 259

Anselm Küsters *
Laura Volkind **
Andreas Wagner ***

Digital Humanities and the State of Legal History. A Text Mining Perspective

* Max Planck Institute for European Legal History, Frankfurt am Main, kuesters@rg.mpg.de
** Instituto de Investigaciones de Historia del Derecho (INHIDE) / Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires
*** Max Planck Institute for European Legal History, Frankfurt am Main

Anselm Küsters, Laura Volkind, Andreas Wagner

Digital Humanities and the State of Legal History. A Text Mining Perspective

Introduction

For reasons of curiosity, we perused the two recent Oxford handbooks on legal history looking for discussions of digital methods in legal history. One of the fundamental decisions to be made when organizing such a handbook is defining which methodological approaches deserve an article of their own and which ones are to be understood rather as cross-cutting themes to be discussed in the context of many articles dedicated to other things. In the case of digital methods in legal history, this decision seems to have been a tough one – at one point, you can find a curious reference to a »chapter on ›Legal History and Digital Humanities‹« (OHBLH 354), but in the final publication there is no such text.

However, discussing digital methods in the context of other subjects has, in our opinion, the disadvantage that more systematic, methodological arguments cannot really be developed. Put more concretely, the most ›substantial‹ contributions regarding digital methods are, for whatever reason, those on »The Intellectual History of Law« by Assaf Likhovski, on »Taking the Long View« by Paul D. Halliday, on »Quantitative Legal History« by Daniel Klerman, and on »Indian Law« by Mitra Sharafi, all of which are in the *Oxford Handbook on Legal History*. (Equally surprising, there is no mention of digital methods at all in Angela Fernandez's »Legal History as The History of Legal Texts«.) However, even these articles do not really ›discuss‹ digital methods, rather they merely refer to them (and to some projects) as contributions of sorts to their respective fields of interest.

Thus, if you are looking for digital methods in those handbooks, you can hardly find more than some namedropping passages where things like »digital mapping […], network analysis […], text analysis« (OHBLH 845f.) are mentioned, together with references to example projects where they have been employed but without any explanation as to:

- why these methods are mentioned and not others,
- what they are doing, to which end and under what circumstances,
- what, possibly transformative, impact these methods have on the (respective sub-) field of legal history, and
- what a scholar considering to apply these methods should be aware of.

While the space for this is limited, the present *Forum* contribution tries to mitigate the scarcity of such discussions by presenting and discussing a few textual analyses that make use – for demonstration purposes – of digital methods. Some other methods of analysis, network analysis, and geo-mapping (among others), cannot be covered here, but we provide a link to an online bibliography where you can find them applied to legal history or a related domain, and discussed critically. A general discussion of digital perspectives beyond concrete methods of analysis concludes this contribution.

Exemplary Analyses

Legal history is concerned with texts to an even greater extent than humanities in general. Through writing, norms achieve stability and communicability, and the vast majority of research in legal history deals with text. Therefore, in our exemplary analyses, we are focusing on a set of methods of textual analysis. More specifically, we will present an analysis using Structural Topic Modeling, followed by an analysis that further investigates one hypothesis resulting from this Topic Model in a corpus linguistics workbench called *TXM*.

Corpus Preparation

First of all, we have prepared all contributions to the two handbooks as a corpus: We have scraped (via copy-and-paste in the web browser) the plaintext from 107 articles via OUP's *Oxford Handbooks Online* site[1] and saved them as ›.txt‹ files (including notes and references, but without abstracts and keywords). Also, we have established a spreadsheet file (in ›.csv‹ format) with title, author, name of the corresponding plaintext file, and the following

metadata fields for each contribution: how many authors the contribution has; their sex, affiliation, place and country of the affiliated institution; which of the two books the contribution features in; the DOI for the contribution, keywords, and abstract. This constituted a corpus of roughly 1,235,000 words (called ›tokens‹) formed out of a vocabulary of roughly 45,000 different basic words (or ›lemmata‹).[2]

Topic Modeling (STM)

Besides more general labels like ›text-mining‹ or ›network analyses‹, Topic Modeling is mentioned explicitly as a method in the handbooks (in Paul D. Halliday's »Legal History: Taking The Long View«, OHBLH 338), and we decided to use this method to illustrate some of the possibilities of quantitative Text Mining. Thus, we used the R language's *stm* package to apply a so-called Structural Topic Model (STM) to the two Oxford handbooks.[3] This technique enables researchers to discover topics within a larger collection of texts and to estimate their relationship to document metadata.

But what exactly is a topic? Topic models treat topics as probability distributions over words, meaning that the estimated model returns several lists of words that have been identified computationally as having a high probability of occurring together. Anticipating our results, figure 1 presents an example for such a list as inferred from the two handbooks. It consists of words such as *genocide*, *nazi*, *jewish*, *criminal*, and *tribunal*,[4] which suggests that the topic encompasses the discourse on *National Socialism (NS) and Law* that is present in many handbook articles (e. g. Randall Lesaffer, »The Birth of European Legal History«; Michael Stolleis, »European Twentieth-Century Dictatorship and the Law«). The topic is displayed as a word cloud, which is a popular way of presenting Topic Modeling output.

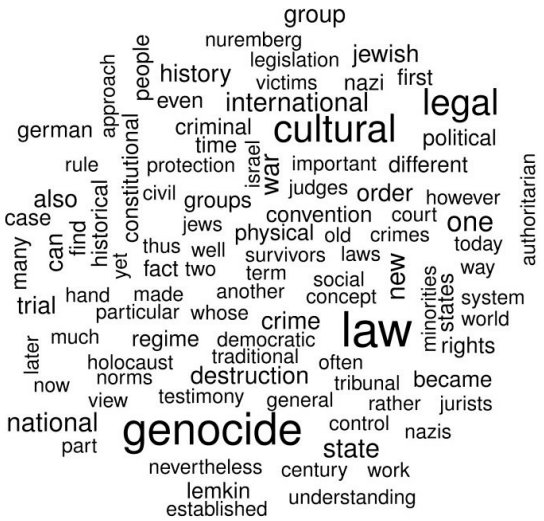

Figure 1

In order to estimate a meaningful STM, that is a set of such lists, we followed a trial-and-error process based on statistically-derived suggestions provided by the software. To determine the optimal topics number, one should test different models and consider the results in terms of interpretability with regards to the specific research question, and then possibly diverge from the merely statistical ›optimum‹. In the end, we opted for a 20-topic model with the estimated topics being displayed in the table presented in figure 2.

1 https://dx.doi.org/10.1093/oxfordhb/9780198794356.001.0001and https://dx.doi.org/10.1093/oxfordhb/9780198785521.001.0001. At this point, credit should be given to Oxford Scholarship Online generously supporting Text and Data Mining for non-commercial purposes (cf. https://www.oxfordscholarship.com/page/FAQs_oso/frequently-asked-questions-faqs#TDM; all links have been last checked on 19 July 2019).

2 For copyright reasons, we obviously cannot publish the full corpus, but we have put the metadata spreadsheet online at https://owncloud.gwdg.de/index.php/s/NTzFsPeFlU3AUVc.

3 Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoldi, The Structural Topic Model and Applied Social Science, in: Advances in Neural Information Processing Systems (NIPS) 26 (2013), paper prepared for the Workshop on Topic Models: Computation, Application, and Evaluation, https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf; cf. also http://www.structuraltopicmodel.com/.

4 Within the framework of Topic Modeling, it is common practice to highlight the individual words (tokens), which are contained in the corpus, in lower case and in italics. In the later part on TXM, the tokens are also given in italics, but not always written in lower case, since their original spelling was retained for the TXM analysis.

STM Output	Label
Topic 1: biannual, contextualize, curricula, dictionary, post-second, non-western, paper	Legal Scholarship in the 20th Century
Topic 2: topical, ancient, unquestioned, decidendi, ellesmere, deviating, historicization	History of Legal Ideas
Topic 3: creoles, spaniards, pre-conquest, conquest, cabildos, hispanic, burgos	Spanish Law and Colonisation
Topic 4: abundance, strata, orality, muslim, scriptural, reliability, matched	Scriptural Law
Topic 5: byzantines, justinianic, gaian, imperial, imperial, convenience, applicability	Roman Law
Topic 6: recension, concordance, modicum, sinners, sacraments, sinner, fournier	Canon Law
Topic 7: systèmes, grands, inter-state, international, comparatist, emer, vattel	Comparative Law
Topic 8: law – public, lettre, forests, health, portray, earth, rivers	Environmental Law
Topic 9: römische, mid-eighteenth, mid-eighteenth-century, theory, pride, weberian, introductory	History of Legal History
Topic 10: trials, jury, murders, adversary, negotiating, fined, indictment	Criminal Law
Topic 11: owns, wild, futile, acres, hunt, hunting, filed	Agricultural Property Law
Topic 12: parties, empirically, dissolution, recommendations, marketplace, economists, apogee	Economic Legal History
Topic 13: coincidentally, prehistory, connect, fruitful, intensely, song, laypersons	Textual Analysis
Topic 14: buttressed, undergoing, outstanding, ports, advocate, hearings, falling	Civil Law Procedures of Juridical Hearings
Topic 15: worker, producers, centrally, businesses, observers, graduates, towering	Marxist Legal History
Topic 16: panels, spent, elimination, ipso, judges, appealing, procedurally	Common Law Procedures of Juridical Hearings
Topic 17: adenauer, gaulle, technocratic, reuter, decisional, dual, knew	EU Legal History
Topic 18: jurisprudence, championed, formalists, happy, formalist, self-interest, dictate	Natural Law vs. Formalism
Topic 19: quantities, folios, inclined, possesses, useless, remarked, grants	Method of Legal History
Topic 20: adolf, eichmann, immunity, nazis, persecution, israel, testimonies	NS and Law

Note: For each estimated topic, the table gives a list of the seven most important words (i.e. the words with the highest probability of being named within that topic) as well as the manually added label. The seven words are ranked by statistical importance. The specified words given in this table are manually cleared word forms of the underlying tokens. Since no lemmatization procedure was applied when creating the corpus, the latter contains the actual word forms as used in the handbook articles, including apostrophes, quotation marks, or punctuation marks. These characters, which have only a grammatical function, have been manually removed for the table to ensure better readability. For example, parties' *was shortened to* parties*.*

Figure 2

As the STM produces groups of words that merely have a high probability of occurring together, topics are usually referenced by their respective top-scoring words (according to various measures such as intra-group probability, distinctiveness vis-à-vis the other groups, etc).

Since the actual reason underlying the groups' respective coherence is unknown to the STM, the researcher normally also assigns labels to the groups, as done in the right column of the table above. Usually, topics evoke specific associations, so that reasonable and coherent labels can be inferred relatively quickly. We give two examples. The seven most probable words for Topic 12 include *empirically*, *marketplace*, and *economists*, which clearly signals a proximity to *Economic Legal History*,[5] as,

5 Just as tokens are marked in a certain way (lower case, in italics) in a Topic Modeling analysis, topic labels are highlighted in the text in italics, but in capital letters.

for instance, discussed in the articles by Ron Harris (»The History and Historical Stance of Law and Economics«) and Anne Fleming (»Legal History as Economic History«). For Topic 17, we can identify names such as *adenauer* and *gaulle* and terms like *technocratic*, which can be linked to *EU Legal History*, and are in turn reviewed in the two articles of Peter Lindseth (»Foundings: European integration«; »The Law of the European Union in Historical Perspective«).

However, topics are not always recognizable at first sight. If a topic lacks a straightforward interpretation, it is helpful to read the texts that exhibit a large share of this topic in order to get a better sense for the proper interpretation of the word list and thus the appropriate label. This procedure had to be followed for most topics in the table above, since the specialized vocabulary and the wide topical variety made it relatively difficult to find intuitive common denominators.

Finally, a well-known fact in Topic Modeling (and yet a common source of misunderstandings and criticism) is that topics do not necessarily have to describe a straightforward theme, in the sense of a subject matter, but that they can also form clusters of methodological words, days of weeks, person's names, or rhetorical devices. In our example, this happened in the case of Topic 13, which features many rhetorical terms (*coincidentally*, *connect*) and even metaphorical words (e.g. swan *song*, siren *song*) that were utilized in diverse articles, irrespective of the particular theme discussed. While scholars commonly use labels like *Descriptive Language* or *Rhetorical Elements* when dealing with such topics, we opt for the label *Textual Analysis* because the manual revisiting of the corpus and close reading revealed that the specific terms listed as Topic 13 often appear when scholars discuss their own (or others') textual analysis of certain sources (e.g. source X is particularly *fruitful* for the question of Y; X was found to be a particularly *fruitful* concept when analysing Y; studies on X have concerned themselves *intensely* with Y). Thus, Topic 13 should not be interpreted as reflecting textual analysis method or textual analysis as such, but as reflecting the rhetorical expressions frequently used when summarizing the results of such analyses. Note that, generally, the STM found all 20 topics without knowing that it deals with a set of legal historical articles and without any pre-coded definitions or lists of key terms. Yet it came to results that correspond, to a large extent, to the semantic and contextual meaning that the words actually exhibit in the corpus (e.g. sorting *vattel*, *adenauer*, and *eichmann* to different topics [7, 17 and 20], but *adenauer*, [de] *gaulle* and even [Paul] *reuter* to the same topic [17]).

Besides inferring topical content, Topic Modeling allows us to structure large quantities of texts by providing different means of corpus level visualization. The most popular one relates to the expected proportion of the corpus that belongs to each topic. This is plotted for the estimated STM in figure 3. We see, for example, that the *NS and Law* topic (20) introduced in the beginning is actually a relatively minor proportion of the overall legal-historical discourse. The most common topics refer to *Roman Law* (5), to a general topic full of words that historians commonly utilize for reporting about *Textual Analysis* (13), and – not surprisingly for handbooks that intend to present the evolution of a discipline and its state-of-the-art – to a topic on the *History of Legal History* (9).

We now discuss estimating topic-metadata relationships, as the ability to plot these relationships is the key benefit of STMs. This feature has been used in the social science literature to model, for instance, the framing of international newspapers, Twitter feeds, and religious statements.[6] There are two ways in which the metadata can enter into our model: Whereas in topical prevalence, the metadata values of the various documents affect the frequency with which a topic is discussed in the respective document, in topical content, they influence the word probability distribution ›within‹ a specific topic in a document. In this example, we use the handbook variable (OHBLH vs. OHBELH) and the author's country as covariates in the topic prevalence portion of the model and the handbook variable again in the content portion.

First, we would like to plot the change in topic proportion shifting from one handbook to the other. Since our covariate of interest is binary, we estimate the expected proportion of an article that belongs to a topic as a function of a first difference type estimate, where topic prevalence

6 The authors of the *stm* package provide a list of articles using STM at their website mentioned above.

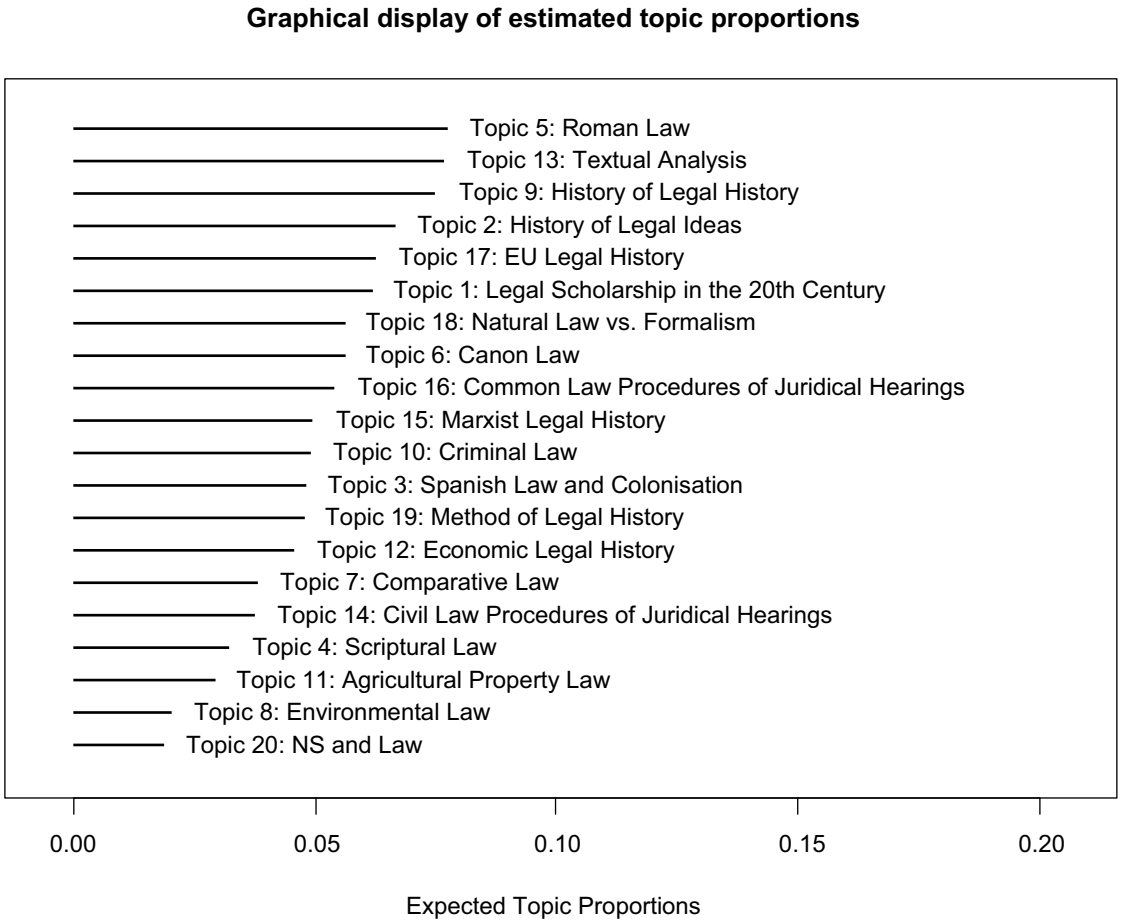

Figure 3

for each topic is contrasted for these two groups (OHBLH vs. OBHELH). Figure 4 gives the results. We see that *Legal Scholarship in the 20th century*, *Comparative Law*, *Textual Analysis*, and *Natural Law vs. Formalism* are strongly discussed in the contributions to the *OHBELH*, while topics on *Canon Law*, *Criminal Law*, and *Method of Legal History* were largely associated with writers for the *OHBLH*.

We can use the same method to investigate changes in topic proportion associated with the authors' countries of residence, since this information was also included as a covariate in the estimation of the STM. To give an example, we contrast authors that are located in the US with authors affiliated with German institutions. Inspecting the corpus reveals that, overall, there are 33 US-based authors and 14 Germany-based authors that have published articles in the two handbooks. When plotting topic prevalence for all 20 topics given in these two groups, it becomes clear that the country of residence has indeed some significant correlation to the author's choice of topics (fig. 5). US-based authors are more likely to write about *Roman Law*, *Comparative Law* and *Natural Law vs. Formalism*, whereas authors based in Germany tend to write on *Canon Law*, *Economic Legal History*, *Marxist Legal History* and *EU Legal History*. It should be noted, however, that these effects only indicate statistical correlations, not causations. For example, the authors might be writing about a certain subject mainly because the handbook editors have asked them to do so rather than because of the location of their institutional affiliation. Moreover, the relatively small sample size of our handbook corpus (typical Topic Modeling projects cover millions of tokens) increases the likelihood of sample selection bias.

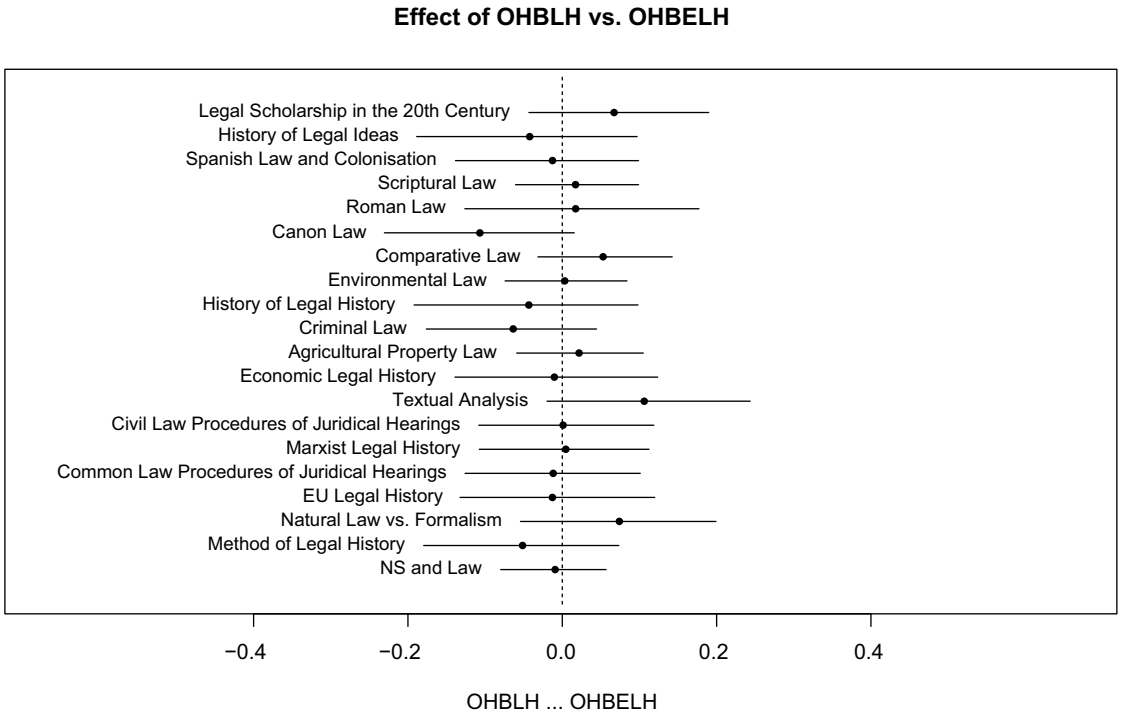

Figure 4

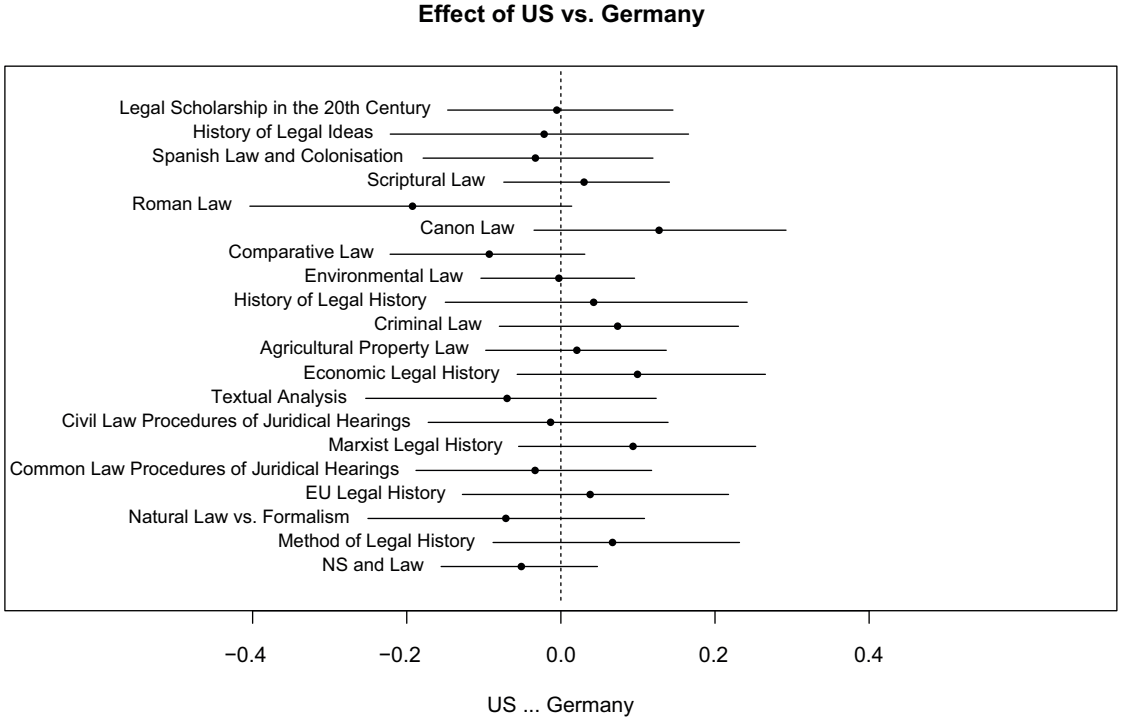

Figure 5

Finally, we can analyze the influence of the respective handbook as a topical content covariate. This allows us to investigate which words ›within‹ a certain topic are more associated with one handbook versus the other. In our analysis (not shown here), we plotted vocabulary differences by handbook for the *NS and Law* topic (20), whose top seven words as displayed in the general table are *adolf*, *eichmann*, *immunity*, *nazis*, *persecution*, *israel*, *testimonies*. However, as calculations make clear, the two handbooks treated this topic very differently. In particular, authors of the OHBELH were much more likely to use words such as *state*, *national* and *german* when writing about *NS and Law* (20), whereas OHBLH authors emphasized terms such as *genocide* and *cultural*. There might be an intuitive explanation for this: Whereas a volume that focuses on European legal history might be more inclined to refer to classic national histories of states and to their respective laws, a handbook trying to provide a global perspective on legal history is more likely to draw on aggregating meta-concepts like *genocide* and *culture* when referring to the legal system of the Third Reich. (In actual fact, something else is going on here – a factor related to the small sample size and that will be discussed in the next section.)

But first let us acknowledge that estimating a Topic Model, such as the STM discussed in this review, has three important benefits not easily achievable by means of the classic close reading of texts: First, this method does not require the imposition of pre-defined categories and is thus somewhat shielded from bias – or at least, it isolates and makes more explicit the introduction of a schema of interpretation by the researcher. Second, topics are explicit, so other researchers can reproduce the analysis or challenge the labels associated with the topics. Third, the computational power allows us to understand and structure corpuses of texts that are difficult to grasp coherently for a single scholar due to their length. This might not be entirely true for the two handbooks analysed here, which ›only‹ encompass 2,374 pages, but it becomes much more relevant when dealing with, for instance, a large historical newspaper archive. Nevertheless, as has become clear as well, these quantitative techniques still depend on the researcher's judgment. They may serve as exploratory tools that stimulate new questions and hypotheses to be tested or complement – and not substitute – existing tools of legal historical research.

Corpus Linguistics (TXM)

Topic Modeling is a relatively recent method, and it is one in which many things are being accomplished without the assistance of the researcher. While this reduces chances of introducing bias, it also makes it harder for the researcher to provide interpretations or to avoid over-interpretation when she may be ignorant of all the steps involved.

Therefore, we also want to present a more ›conventional‹ analysis of our OHB corpus using various functions of a powerful corpus linguistics platform. Corpus linguistics workbenches, or toolkits, like *GATE*, *TXM* or *WebLicht* allow the researcher to quickly gather statistics about aspects of language use in the assembled corpus.[7] Basically, one can see specific word forms or basic words ranked by their frequency (fig. 6). For what it is worth, the most frequent basic word in our corpus, *the*, comprising its specific forms *the* and *The*, occurs 73,149 times. The next most frequent words are *of*, *and*, *in* and the various forms of the verb *be*, all of them being so-called function words. The high frequencies of the content words *law*, *legal*, and *history* are also hardly surprising.

In all likelihood, content words related to specific research questions are more interesting, but then of course it depends on the researcher's creativity and experience to translate his or her research question into query terms. Suppose the respective weight of justice and power is at issue. We can use TXM's ›index‹ and ›progression‹ tools to see that both terms cumulate more or less constantly over all the articles, but that the curve for *power* is more even and steeper, and that it totals at almost double the frequency of *justice* (1,164 vs 552 occurrences).

A central function of corpus linguistics is the creation and contrasting analysis of sub-corpora. TXM allows us to create sub-corpora (a corpus

7 For GATE, see https://gate.ac.uk/; for TXM, see http://textometrie.ens-lyon.fr/spip.php?article67&lang=en; for WebLicht, see https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page; also, you might have a look at the better-known and easier to use, but in some ways less flexible Voyant Tools at https://voyant-tools.org/.

Lemma	freq	Lemma	freq	Lemma	freq	Lemma	freq	Lemma	freq
the	73149	)	25020	for	7230	history	5158	their	3042
,	69520	(	24955	legal	7156	with	4851	its	2938
.	51231	to	21575	:	7012	or	4311	Law	2886
of	48918	’	21271	have	6794	from	4292	but	2840
@card@	44073	a	17006	this	6309	not	4252	at	2708
and	30489	law	12478	on	6073	which	3680	they	2291
in	26231	as	9637	by	5912	;	3580	also	2267
be	25959	that	9608	it	5688	an	3425		

Figure 6: Most frequent lemmata

being just a part of the full corpus) and partitions (a non-overlapping, collectively exhaustive set of sub-corpora) according to the metadata values that we have recorded beforehand. One could, for instance, partition by authors' sexes, and contrast, e. g. the mere number of words written by women (269,218) to those written by men (967,440; this would be even more dramatic when applied to the European handbook alone: 53,187 vs 577,862).

Alternatively, one could partition the corpus according to the country that the author's affiliation is located in, or according to the affiliation itself, and again report the number of words per partition (fig. 7).[8]

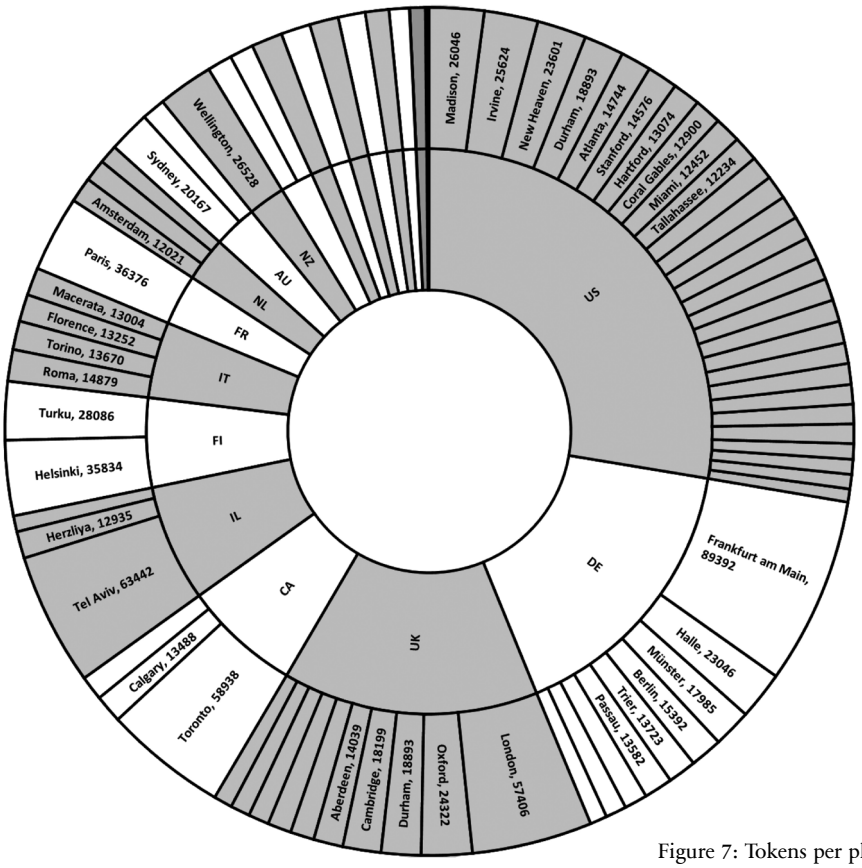

Figure 7: Tokens per place

8 The image in figure 7 contains slices per country and per location, sized proportionally to the respective number of words / tokens in the corpus. The labels of the slices are either the country code or the place that the author's respective affiliated institution is located, plus the number of tokens from this place. In cases where this information did not fit into the slice, there is no label.

Or, to enter a bit deeper into the linguistic aspect, one could contrast the partitions' vocabulary content rather than their mere size. TXM calculates a ›specificity score‹ for each word, based on the deviation of the actual from the expected number of its occurrences in a partition (given the partition size and the total number of occurrences in the whole corpus).[9] In this way, researchers can gain another perspective on the contrast between the two handbooks.

Among the words specific to the European handbook (see also fig. 8), we see:
- named entities, in particular the names of European nations (like *France*, *Denmark*, *Sweden*, but also as adjectives – *German* – and referring to historical entities *Roman* and *Byzantine*),
- function words in other European languages that probably come from literature in those languages being cited (*und*, *de*, *der*, *des*, *die*, *im*, *et*, *zur*), and also
- some words that seem to indicate subject matters more prominent in the European handbook than in the ›global‹ one (*royal*, *king*, *church*, *kingdom*, but also *court*, *city*, and *town*).

In the list of words specific to the ›global‹ handbook, by contrast, the perspectives that seem to suggest themselves are (see also fig. 9):
- very general (first and foremost *history*, *historian* and *historical*, *past*, or *jurisprudence*, *research*, and *scholarship*) and
- methodological (the general *analysis* and *inquiry*, but also *critical*, *realist / realism*, and *feminist*), but there are also
- some terms indicating concrete subject matters or fields of law (*Islamic*, *environmental*, *violence*, *Jewish*, possibly *black*).

But let us come back to our *NS and Law* topic from the preceding section. For a more detailed assessment, we have queried 9 terms related to crimes against humanity (*genocide*, *torture*, *deportation*, *displacement*, *rape*, *enslavement*, *persecution*, *cleansing*, *massacre*) and a further 5 terms related to German National Socialism (*NS*, *NSDAP*, *Nazi*, *Nazis*, *Nazism*). We find that 7 of the 14 terms occur more than 10 times in the two handbooks. Looking up the specificity values of these 7 terms for some of the countries of the corpus' authors, the picture shown in figure 10 emerges.

It is perhaps worth noting that there is a so-called ›banality‹ threshold within which fluctuations of usage of the terms are not really significant, and we have left this threshold at the default value (of ± 2.0, indicated by thin lines in the figure). We see that UK- / US-based authors seem to avoid all the terms mentioned to a non-trivial degree; arguably, they do not treat the topic to any extent at all. Moreover, Australian and Finnish authors conspicuously refer exclusively to *rape / displacement* and, respectively, to *torture*, which none of the others seems to touch upon. This fact might indicate that it was (most likely) misleading to approach the topic solely from the perspective of crimes against humanity, assuming that many of the terms would typically occur together, which, if true, could have been motivated by this legal concept.

Anyway, at least the numbers seem to confirm that German authors discuss the topic using the term *NS*, whereas Israeli authors rather use *genocide* and *Nazi / Nazis*. However, here we encounter again problems connected with the small sample size and selection bias alluded to above. Building a sub-corpus for only Israeli authors, partitioning that sub-corpus according to author, and then revisiting our topic's terms, we find that it is in fact only one single contribution that produces the particular profile of the ›Israeli way‹ of discussing the topic and using the vocabulary of *genocide*; an unsurprising result given the contribution's title: »Cultural Genocide: Between Law and History« by Leora Bilsky and Rachel Klagsbrun. It is quite likely that this even spills over and produces the would-be ›OHBLH way‹ of discussing it. And vice-versa, just one single contribution (Michael Stolleis' »European Twentieth-Century Dictatorship and the Law«) is responsible for the ›German‹ (and for the ›OHBELH‹) way of discussing the topic, mentioning terms such as *NS* more than

9 The mathematics behind TXM have been discussed in Pierre Lafon, Sur la variabilité de la fréquence des formes dans un corpus, in: Mots 1 (1980) 127–165, https://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008.

Lemma	freq	total	score
und	730	799	134,5924
Roman	1497	1913	133,9929
de	936	1123	114,007
der	595	651	110,059
des	561	629	93,1504
@card@	24553	44073	93,0246
royal	434	462	91,0955
European	1056	1368	88,479
the	39559	73149	69,191
die	421	474	68,8671
Europe	616	755	68,2762
king	317	339	65,3845
court	1232	1756	59,7606
im	252	262	59,3965
Magdeburg	200	200	58,6269
ius	266	283	56,2841
city	308	341	54,4902
century	1429	2112	54,1823
town	252	268	53,4559
Byzantine	164	165	46,1619
territory	246	271	44,6531
justice	494	643	40,9614
church	219	240	40,7571
et	399	502	39,3786
ecclesiastical	219	242	39,3089
Code	280	327	39,2609
Recht	213	234	39,238
zur	169	177	38,681
Church	220	246	37,505
German	573	779	37,5046
France	304	366	37,1181
iuris	174	186	36,2849
Denmark	129	130	36,0044
medieval	443	580	35,7486
Scandinavian	113	113	33,1225
territorial	176	193	32,8112
Sweden	147	155	32,6996
Ages	204	233	31,7762
kingdom	123	126	31,4576
droit	227	267	31,0049
lord	156	169	30,7441
Italy	143	152	30,6975
Böhlau	104	104	30,4843
emperor	175	195	30,4481
jurisdiction	412	549	30,384
)	13637	25020	29,9753
Danish	102	102	29,898
n	1179	1842	29,195
Scottish	118	122	28,8655

Figure 8: Most characteristic lemmata in OHBELH

Lemma	freq	total	score
ff	1718	1950	291,9503
history	3527	5158	172,8805
historian	903	1079	123,7689
historical	1271	1646	120,8141
American	798	936	118,5407
L	554	625	97,2601
ibidem	374	387	95,6685
that	5690	9608	88,5606
Islamic	335	352	79,9611
U	282	289	75,7315
analysis	531	641	70,0759
what	986	1367	66,6389
past	432	508	63,9505
how	721	951	63,206
we	923	1301	57,0506
legal	4184	7156	56,8638
critical	384	452	56,7178
S	344	403	51,9324
law's	264	291	51,5362
.	26814	51231	50,9613
:	4060	7012	49,3327
environmental	169	171	48,6632
realist	171	175	46,4482
feminist	145	146	42,9503
’	11432	21271	42,4641
inquiry	188	202	41,0549
scholarship	377	481	39,1924
research	392	505	38,9399
violence	241	279	38,502
Jewish	175	188	38,2916
critique	215	243	38,0036
black	143	150	34,7683
r	303	381	33,8323
Holmes	151	162	33,319
New	430	586	32,4286
jurisprudence	425	578	32,3995
archive	124	129	31,3408
realism	126	132	30,8727
Legal	730	1110	28,5839
?	628	935	28,3649
study	758	1165	27,5411
gender	121	129	27,538
think	316	420	27,2591
economics	135	149	26,5801
work	937	1490	26,2551
Bentham	156	179	26,138
our	339	462	25,7715
queer	82	82	25,348
Tomlins	116	125	25,2789

Figure 9: Most characteristic lemmata in OHBLH

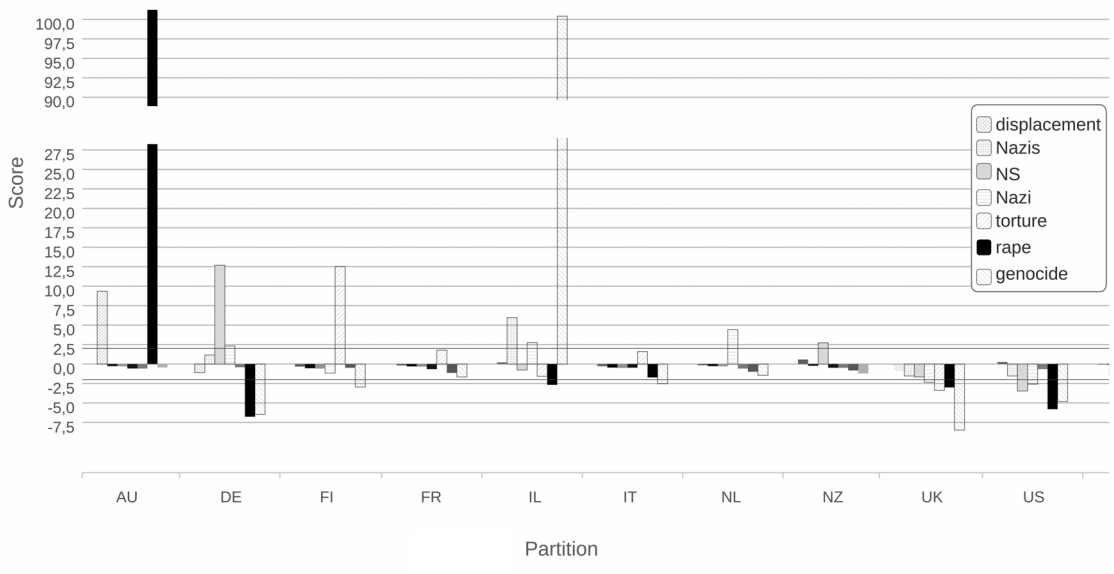

Figure 10: Specificity scores for »NS and Law« terms

others. So it is certainly mistaken to infer from them either a rhetoric that would be characteristic to some extent for all authors of a certain national tradition or some preference in the respective editors' policy of inviting contributors that would adhere or not to a certain rhetoric! And whether the particular profiles of the two relevant contributions resulted from the chosen or requested topic, from developments that the authors may be involved in on their respective national level, or from the authors' idiosyncrasies cannot be decided by corpus linguistic means.

Thus, one of the key takeaways is that relating findings of digital methods to research questions is something that requires scholarly interpretation, contextual knowledge, and close reading of the respective documents. (On the other hand, this makes the fact that STM was nevertheless capable of sorting the terms *genocide* and *NS* into the same topic in the first place all the more interesting.)

Another key takeaway might be the following: Both Topic Modeling and more conventional corpus linguistics are most useful when assessing discourses instead of opinions or statements. The researcher's goal in using these methods should not be to understand what individual documents assert without reading them; nevertheless, such an approach could more plausibly be used to learn about various ways of talking and writing more easily discerned in large sections of a given discourse. Once made visible, it then becomes possible to interpret and reflect about how these ways of talking and writing might frame certain subjects.

With this in mind, we want to focus on more cross-cutting phenomena and offer a final example for this approach. As we have seen, the contrast between *power* and *justice* is ubiquitous and further investigation warranted. However, it would probably be more fruitful to return to the question posed at the very outset: How well established are digital methods and resources within the discipline? First of all, we can see that there is a steady occurrence of references to online resources (by *http(s)* or, less frequently, by *doi*), resulting in at least 225 references to online resources.

Then, we can have TXM list all words that occur together with any word beginning with *digit* (in a ›window‹ of 20 words to the left and 20 words to the right). The most significant co-occurrent is *humanity*, certainly because ›digital humanities‹ is an established (and fashionable) term. Co-occurrents like *opportunity* (score: 5.3), *possibility* (2.7), *access* or *accessible* (5.7/2.4), *available* (5.8), and *use* (6.3) suggest that, if things digital are discussed, the attitude seems to be rather open and optimistic and there seems to be a certain focus on the ways in which resources are available in digital form. This last point is reaffirmed by the prominence of co-occurrents like *archive(s)*, *source*, *database*, *digitization*, *manuscript*, *newspaper*, *library*, *collection*. Terms that might indicate a more skeptical attitude like *issue*, *miss*, *serious* seem to do so only in

Coocc	Freq	CoFreq	Score	MeanDist	Coocc	Freq	CoFreq	Score	MeanDist
digital	51	15	25,431	12,133	for	7230	48	6,250	8,563
humanity	57	14	22,530	4,357	Manifesto	23	4	6,174	15,000
archives	74	14	20,777	12,429	oral	111	6	5,914	8,500
tool	151	15	17,799	3,933	available	183	7	5,794	4,429
history	5158	58	16,246	9,103	access	189	7	5,701	16,000
source	870	25	16,065	9,080	search	122	6	5,675	10,167
/	1344	29	15,268	11,862	opportunity	81	5	5,301	6,200
Digital	21	8	14,912	11,500	Armitage	38	4	5,269	14,250
India	150	13	14,739	9,846	digitize	13	3	5,129	14,000
database	16	7	13,631	9,286	Cast	2	2	5,051	5,000
digitization	20	6	10,580	13,167	Doctoral	2	2	5,051	9,000
paper	43	7	10,212	8,000	Enough	2	2	5,051	5,000
Indian	159	10	10,104	7,100	Nystrom	2	2	5,051	5,000
manuscript	115	9	10,008	7,778	Putnam	2	2	5,051	8,000
Naoroji	6	4	8,928	4,250	Sidonie	2	2	5,051	18,000
Patel	6	4	8,928	14,000	Tanenhaus	2	2	5,051	5,500
newspaper	17	5	8,849	4,000	Text-Searchable	2	2	5,051	1,000
new	1291	19	7,588	8,316	Trove	2	2	5,051	8,500
datum	36	5	7,084	10,400	Good	17	3	4,757	9,333
Dinyar	4	3	6,975	14,000	<	184	6	4,654	9,667
>	185	8	6,943	13,375	Dadabhai	3	2	4,574	5,500
archive	129	7	6,817	4,000	Lara	3	2	4,574	9,000
Library	17	4	6,739	9,000	visualization	3	2	4,574	10,500
collection	225	8	6,297	4,125	visualize	3	2	4,574	7,500
use	1272	17	6,289	6,412					

Figure 11: Co-occurrences of digit

one instance. Figure 11 shows how we can see the immediate context of the respective occurrences in the list of concordances (bottom third of the image); furthermore, it shows how we can then select a passage (line 3, with *digital* being followed by *miss* after five words) and go back to the full text and read the passage in question in full (topmost third of the image). Here we see that it is Paul Halliday discussing the danger of ignoring sources like manuscripts that are not available in digital form merely for this reason.

However, while both aspects – methods and resources – related to the digitization of legal history are represented in the handbooks, only the latter is featured prominently. Fifteen different authors (out of 100 in total) mention some aspect of digital research, and eight do so more than twice. But as we have seen, *archives*, *collections*, or *databases* occur quite frequently in the context of *digit**, whereas references to *digital tools* or *software* are scarce. Only five authors (Likhovski, Halliday, Klerman, Sharafi, the four authors mentioned at the very outset of this review, plus Dirk Heirbaut in the OHBELH) mention these. Assaf Likhovski suggests that the most promising aspect of what he terms the digital revolution is not »the use of new tools to mine this data, but more modest projects: the creation of databases« that help to visualize data and the creation of new, curated, and interlinked teaching tools (OHBLH 160).

However, given that the contributions to the handbooks do not indicate more than a handful of methods, not to mention that in many cases the authors merely refer to the special issue [10] on digital legal history of the *Law & History Review* (2016), more should be done to address such deficits.

10 This is why *issue* has a high co-occurrence score with *digit**, by the way.

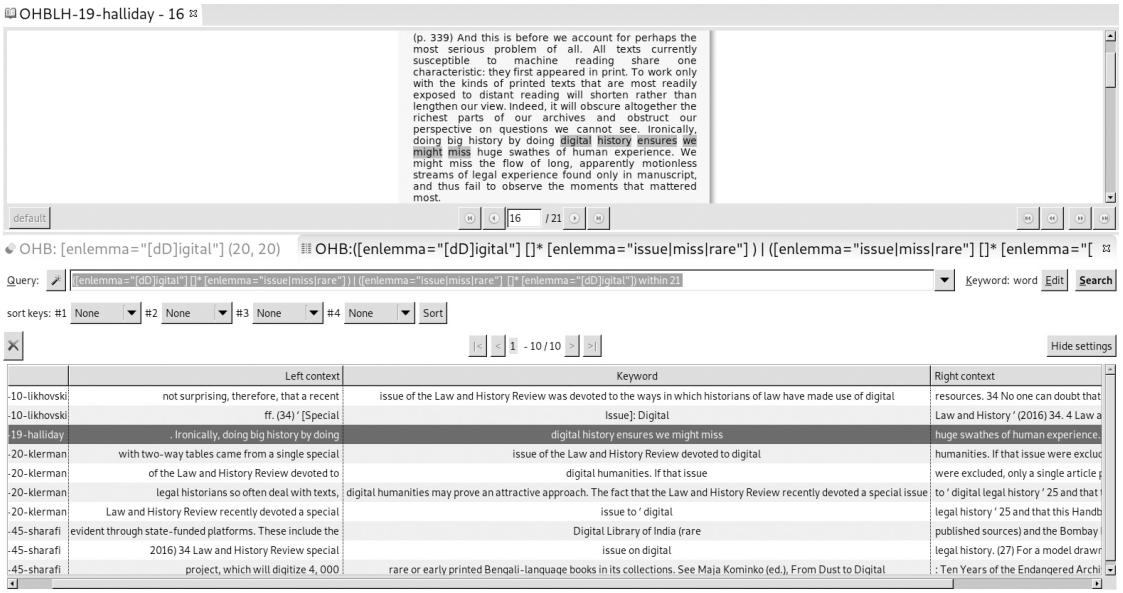

Figure 12: Edition (top) and concordance (bottom) views

There is a clear lack of attractive cases employing such methods, a lack of awareness of available methods, and a lack of opportunities to ›translate‹ digital methods and their technical details to lay – i. e. not-so-tech-savvy – scholars.

More Methods

Due to limitations of space, we are unable to discuss and offer examples of the two other methods mentioned in the handbooks: network analysis and geo-mapping. However, we would like to point out that quite a number of other methods might be relevant to legal historians. Digital humanities projects have already put ›Text Reuse Detection‹ or information extraction methods, such as ›Named Entity Recognition‹, to good use. And in the economically dynamic field of applied law, ›big players‹ like Westlaw, LexisNexis, or Bloomberg, as well as countless IT startups are developing their service portfolio and offer (or are researching) methods of citation recognition, argument mining and evaluation, and recommender systems for judges, litigant parties, or lawmakers.

For all of the approaches mentioned above, we have established an online bibliography and are trying to list literature that is applicable to legal history and / or related fields – or at least introduce and discuss this literature critically.[11]

Discussion

Digital Resources

Even with respect to the resource-focused aspect of digitization, a critical discussion is still lacking. When building a digital resource, one has to check the context and profile of other related digital resources, and the selection of data at the very outset should be examined critically. Can the new resource link to other established resources? Is it capable of helping to establish some other resources? How does it participate, if at all, in a process of canonization or counter-canonization?

Understanding ›data as capta‹, according to Johanna Drucker, draws attention to the process of acquisition and recording of data, where decisions about how to ask, what to record, what to

11 The bibliography can be found here: https://www.zotero.org/groups/2163790/digital_legal_history/items/collectionKey/YEKDRSB9.

ignore, and how to normalize must be made. Also, it is here that biases with regards to the relevance of non-canonized perspectives, opinions, and material come into play. With regards to the technical aspect, for instance, under which conditions are OCR techniques applicable and what are their (dis-)advantages? Or, more in terms of scholarly self-understanding, how does a project position itself with regard to crowdsourcing and the contributions of ›citizen scientists‹?

Data modeling is another crucial point to consider and discuss even before starting the analysis. Are you dealing with a text or something else? If it is a text, is ›text‹ the best form in which to record the information for your project? Might tabular, relational, or semi-structured data be more appropriate? Do you normalize values (and if so, do you keep the original values or discard them?)? What kinds of metadata should go along with the records?

Digital Methods

In the following, we present a selection of questions that digital tools and methods should be submitted to once they come into the purview of legal history. (In the presentation of our STM and corpus linguistics examples above, we have at least hinted at how we would respond to some of the questions for those methods.)

Since most methods accept data and additional configuration parameters, it is important to understand and critically reflect on the parameters used. At what point in the process does one feed a researcher's parameters into the method? Which effects are produced by a change in the parameters, and why would one (rightly or erroneously) enter one value rather than another under actual research conditions? Does the method / tool provide for repeated runs with varying parameters? How do you evaluate the quality of the results of different runs?

In many cases, scholars add annotations to their data and it may be desirable to access these at various stages of the process. For instance, is there a standard data format to adhere to while entering the annotations, and is it possible to access, expose, or export intermediary results (e.g. scan images while you are still waiting for OCR or transcriptions)?

For a number of methods, there is a considerable amount of complexity introduced by sophisticated mathematical algorithms, by the mere fact that parts of the process behave probabilistic / contingently, or by the sheer mass or multidimensionality of the data. It is good to know which parts of the process tend to become non-transparent, and why. Is one able to understand what the algorithm is doing – both in general and more specifically? Is it easy to comprehend what the operations performed on the data mean or represent in real life, or why one would want to do this with the specific data at hand?

Finally, is it clear where the more ›objective‹ part of the process ends and where interpretation begins? How do you avoid reading more into your results than the information warrants? If you catch yourself over-interpreting, is it possible to operationalize the interpretation as another hypothesis, so that it can subsequently be checked and eventually be substantiated?

Opportunities of Digitization

While we have mostly pointed out questions that might possibly help to orient a critical discussion of digital methods and resources, we want to close by highlighting the opportunities that digital methods and resources present. As Mitra Sharafi (OHBLH 847), for example, pointed out, new large-scale digitization projects coordinated and funded by national and international consortia seem to piggyback on the technological advances that image acquisition and OCR are making. And the combination of technological advances and political initiatives may mean better chances for digitally preserving endangered cultural heritage, e. g. from small and / or remote archives or libraries. While the serial character of such cataloging and acquisition work is not completely new, the ratio between effort and benefit has shifted significantly. Moreover, the building momentum will hopefully benefit smaller institutions with valuable holdings yet limited funding as well.[12]

12 See, for example the British Library's *Endangered Archives Programme* at https://eap.bl.uk/.

Unlike the situation a few decades ago, once collections are available in digital form, it very often implies that they are internationally – even globally – accessible and communicable. (The words *available* and *accessible* occur 216 times in the OHB corpus, the most frequent co-occurrents being parts of internet addresses like *www*, *http*, *org*, *blogspot*, *thefacultylounge*, *jotwell*, *nytimes*, *washingtonpost*, etc.) Besides the technical infrastructure, this communicability is facilitated by the establishment of international encoding standards like Unicode, RDA, TEI, and CIDOC CRM, which are transparently developed and recognized by cultural heritage institutions worldwide.[13] The main factor limiting the reach of digitized collections at the moment seems to be licensing and paywall arrangements, but sometimes it is also due to a lack of consideration for user diversity.

Various authors in the OHB corpus acknowledge the new possibilities of searching data once it is available as digital full text data. What they have in mind, however, seem to be primarily ›classical‹ full-text searches of documents that previously could not be searched at all. There are (at least) two other important benefits worth mentioning: First, with searches being carried out by computer systems, linguistic and context searches are now possible (i.e. search X in all its grammatical forms, or search X *near* Y). Second, with collections granting access to standardized, machine-readable interfaces, federated searches have also become a reality (i.e. searches that query multiple repositories at the same time via mechanisms like OAI-PMH or SPARQL).

This last point suggests that it will become easier to launch queries, or work with resources more generally, across disciplinary boundaries: Since most of the encoding standards alluded to above are developed independently of any given discipline or research community, the need for capabilities of translating disciplinary terms to those used by the repository standards is on the rise. Once this has been achieved, however, the same query should apply to related databases from other disciplines with relatively few and minor modifications.

The preceding argument about linguistic searches (which are features of repository or of third-party software) suggests that the boundary between methods and resources sometimes seems to blur. Yet, there are important general opportunities related to digital methods as well. Of course, not all questions can be put to a large-scale corpus, but working at very large scales is a way of working that would not be possible without the opportunities that computer processing offers.

Computer processability also means that data can be duplicated, reorganized, and revised without much effort. Thus, the process of scholarly as well as automatic analysis and annotation can be documented in very fine-grained ways. ›Open Science‹ refers to the possibilities (and ambition) to improve the openness, transparency, and reproducibility of research practice as a whole. Things like web annotation services, public collaboration platforms, versioning control systems, lab notebooks, data publication formats, data repositories, and data publication review literature are already available as tools contributing to this endeavor.[14]

The same flexibility and connectedness also enable the accommodation of multiple dimensions and possibly conflicting interpretations of resources without forcing curators and editors to privilege one over the other(s). Instead, it opens the door to providing dynamic ways of presenting information, shifting emphases, and highlighting different interpretations according to the interests and questions that the users may have.

Finally, in the discussion about Structural Topic Modeling, we have seen that one of the main advantages of digital tools is the promotion of what is referred to as serendipity. The new ways of seeing data, patterns, and relations suggested here are not only relevant to the field of legal history as such, but they also may stimulate questions and hypotheses that would otherwise not

13 See the *Unicode Consortium*, https://unicode.org/; the *RDA Steering Committee*, http://www.rda-rsc.org/; the *Text Encoding Initiative Consortium*, https://tei-c.org/ and the *International Committee for Documentation* and its Conceptual Reference Model, http://www.cidoc-crm.org/.

14 Cf. https://cos.io/; https://okfn.org/; https://web.hypothes.is/; https://demo.codimd.org/; https://ethercalc.net/; https://jupyterlab.readthedocs.io/en/stable/; https://zenodo.org/; https://brill.com/view/journals/rdj/rdj-overview.xml. For most of the services just mentioned, there are also other providers available. Moreover, this list is neither exclusive nor a strong endorsement of these services over others.

have occurred to anyone. These questions and hypotheses could then be investigated in novel or traditional ways, but that is another question for another time. Much work in the humanities is still being attributed to a kind of genius, for better or worse, and, just as they push us to make more explicit many other things that we have become used to presupposing or do implicitly, digital methods may very well turn out to organize and consolidate spaces for scholars' creativity, spontaneity, and intuition. Ultimately, it is up to scholars to actively appropriate digital methods accordingly and establish this vision. After all, the goal is not to restrict ourselves to automatically generated and – in the end – more trivial and predictable ways of doing research, but rather to open up more and develop new avenues of analyzing sources.

■