

# Forschungsgeleitete Vermittlung von Datenkompetenz

## Mediendidaktische Aufbereitung von Fallstudien zu Bildungsangeboten

Evgenia Samoilova<sup>1</sup>, Henny Sluyter-Gäthje<sup>2</sup>, Daniil Skorinkin<sup>3</sup>,  
Hannes Schnaitter<sup>4</sup>, Peer Trilcke<sup>5</sup> und Ulrike Lucke<sup>6</sup>

**Abstract:** Der Beitrag präsentiert ein Vorgehen, wie aus konkreten Forschungsfallstudien heraus und eng an diesen entlang mediengestützte Bildungsangebote erstellt werden können. Die Fallstudie “Spanische Grippe” im Bereich der Geisteswissenschaften zeigt auf, wie historische Zeitungsdaten mittels OCR aufbereitet und analysiert werden können, um den Verlauf der Pandemie in der Berliner Presse nachzuvollziehen. Durch den Einsatz des Vier-Komponenten-Instructional-Design-Modells und des Cognitive Apprenticeship-Ansatzes werden bei der didaktischen Aufbereitung die Forschungsfragen und Aufgaben aus realen Forschungsszenarien in den Mittelpunkt der Lernerfahrung gerückt. Interaktive Lehrbücher wurden als Format gewählt und in Jupyter Books implementiert, um eine hohe Kontrolle über die Interaktion mit den Materialien sowie die Möglichkeit, Inhalte dynamisch zu aktualisieren, zu bieten.

**Keywords:** Datenkompetenzen, Fallstudien, interaktive Lehrbücher, Jupyter Books, Vier-Komponenten-Instructional-Design-Modell, Cognitive Apprenticeship, Problemorientiertes Lernen

## 1 Einleitung

Datenkompetenz beinhaltet kritische und wohlüberlegte Sammlung, Verwaltung, Bewertung und Anwendung von Daten, die über kognitive Aspekte hinaus auch affektive Einstellungen umfasst [Ri15; SBH19]. Sie wird als zentrale Fähigkeit in einer globalen Wissensgesellschaft angesehen und ist unerlässlich in der modernen, datenintensiven Wissenschaft [HBK18]. Der Bedarf an Datenkompetenzen wächst in allen wissenschaftlichen

---

<sup>1</sup> Universität Potsdam, Institut für Informatik und Computational Science, An der Bahn 2, 14476 Potsdam, evgenia.samoilova@uni-potsdam.de, <https://orcid.org/0000-0003-3858-901X>

<sup>2</sup> Universität Potsdam, Institut für Germanistik, Am Neuen Palais 10, 14469 Potsdam, sluytergaeth@uni-potsdam.de, <https://orcid.org/0000-0003-2969-3237>

<sup>3</sup> Universität Potsdam, Netzwerk Digitale Geisteswissenschaften, Am Neuen Palais 10, 14469 Potsdam, daniil.skorinkin@uni-potsdam.de, <https://orcid.org/0000-0002-1845-9974>

<sup>4</sup> Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft, Unter den Linden 6, D-10099 Berlin, hannes.schnaitter.1@ibi.hu-berlin.de, <https://orcid.org/0000-0002-1602-6032>

<sup>5</sup> Universität Potsdam, Institut für Germanistik, Am Neuen Palais 10, 14469 Potsdam, trilcke@uni-potsdam.de, <https://orcid.org/0000-0002-1421-4320>

<sup>6</sup> Universität Potsdam, Institut für Informatik und Computational Science, An der Bahn 2, 14476 Potsdam, ulrike.lucke@uni-potsdam.de, <https://orcid.org/0000-0003-4049-8088>

Disziplinen, insbesondere in bislang datenarmen Bereichen wie den Geisteswissenschaften. Auch in den Verwaltungswissenschaften und öffentlichen Verwaltungen bleibt die Situation trotz vieler Open-Data-Initiativen schwierig, da Zugang und Qualität öffentlicher Daten oft begrenzt sind.

Die Vermittlung von Datenkompetenzen erfolgt derzeit häufig über einen Top-down-Ansatz, der auf umfassenden Kompetenzkatalogen basiert. Diese Datenkompetenzframeworks stellen eine strukturierte Grundlage dar, auf der Lehrpläne und Lernziele entwickelt werden. Beispiele für solche Frameworks sind das von Ridsdale et al. [Ri15] entwickelte Modell zur Bildung in Datenkompetenz und das „Future Skills Framework“ des Hochschulforums Digitalisierung [SBH19].

Im deutschsprachigen Raum gibt es einige Angebote zur Vermittlung von Datenkompetenz für Forschende, darunter Universitätskurse, spezielle Programme der Graduiertenkollegs, sowie Online- und Präsenzworkshops und Sommerakademien. Diese sind zwar oft kostenpflichtig, jedoch weitgehend zugänglich. Zusätzlich zu diesen traditionellen Angeboten stellen Plattformen wie zum Beispiel der SSH Open Marketplace [SS24] und Programming Historian [Cr16] wertvolle Online-Ressourcen bereit. Der SSH Open Marketplace bietet eine zentrale Anlaufstelle für Werkzeuge, Dienste und Trainingsmaterialien für die Sozial- und Geisteswissenschaften. Programming Historian veröffentlicht frei zugängliche, begutachtete Lektionen, die digitale Werkzeuge und Methoden erläutern und strukturiert aufbereiten, um eine direkte Anwendung in der Forschungsarbeit zu ermöglichen.

Jedoch verlangen traditionelle Angebote zur Vermittlung von Datenkompetenzen oft ein umfangreiches zeitliches Engagement, das nicht immer mit den Bedarfen und Möglichkeiten der Forschenden übereinstimmt. Online-Formate zur punktuellen Vermittlung von Datenkompetenzen, insbesondere ohne direkte Lehrendenbegleitung, bieten dagegen häufig keine Interaktivität und kein Feedback.

Vor diesem Hintergrund wird im vorliegenden Beispiel ein Vorgehen präsentiert, wie aus konkreten Forschungsfallstudien heraus und eng an diesen entlang mediengestützte Bildungsangebote erstellt werden können, die sowohl relevante Datenkompetenzen in der Breite abdecken als auch die spezifischen Bedarfe konkreter Disziplinen bzw. einzelner Wissenschaftler:innen adressieren. Die Bildungsangebote und deren Evaluation sind noch in der Entwicklungsphase.

## **2 Verzahnung von Forschen und Lernen**

Für den systematischen Aufbau von Datenkompetenzen in der Breite der Fächer müssen die Datenmethoden-liefernden Fächer (wie Informatik oder Informationswissenschaft) enger und systematischer mit den Daten-verarbeitenden Anwendungsfächern verzahnt werden [Lu24]. Ein Beispiel für diese Symbiose sind die sog. Digital Humanities (DH). Der

Bund hat mit der Förderung von Datenkompetenzzentren hier einen wichtigen Impuls gesetzt, der aber mittelfristig in tragfähige Strukturen auch ohne Projektförderung überführt werden muss. Daher braucht es einen Ansatz, der die beteiligten Fächer (d.h. individuelle Forschende, lokale/regionale Institutionen und einschlägige Fachgesellschaften) systematisch miteinander verzahnt. Eine solche Verzahnung findet im Datenkompetenzzentrum QUADRIGA [Bu24] über drei Mechanismen statt:

- In *Forschungsorten* entstehen Fallstudien aus Daten-basierter Forschungstätigkeit, die anhand eng umgrenzter Forschungsfragen die dazu gehörenden Datensätze und Methoden kapseln.
- Diese werden in *Lernorten* zu mediengestützten Bildungsangeboten aufbereitet, die interdisziplinär anschlussfähig sind, als offene Bildungsressourcen bereitgestellt werden und soweit möglich kooperativ bearbeitbar sind.
- Die *Vernetzungsorte* tragen für den Anschluss von entstehenden Fallstudien und daraus resultierenden Bildungsangeboten an einschlägige Fachgesellschaften und Datenkompetenz-Initiativen Sorge.

Damit wird sichergestellt, dass sowohl der aktuelle Stand der jeweiligen Fächer reflektiert ist als auch die bestehenden Kanäle und Formate für den wissenschaftlichen Austausch genutzt werden. Abb. 1 visualisiert dieses Zusammenspiel.

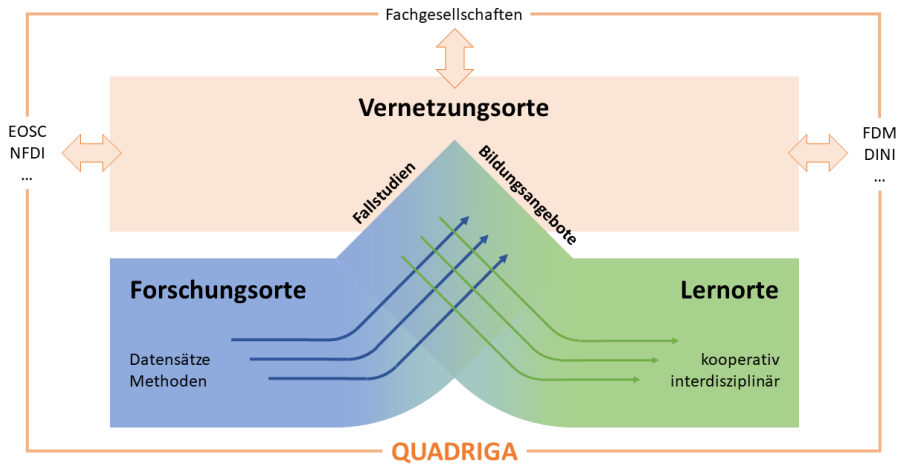


Abb. 1: Verzahnung von Forschungs-, Lern- und Vernetzungsorten entlang von Fallstudien [Bu24]

Die Fallstudien gliedern sich entlang der Datentypen Text, Bild und Tabelle. Sukzessive entstehen in den Forschungsorten neue Fallstudien (aus Anregungen in den Vernetzungsorten), die in den Lernorten aufbereitet und als Bildungsangebote bereitgestellt werden (wiederum über die Vernetzungsorte). Der vorliegende Beitrag präsentiert das Vorgehen für diese mediendidaktische Aufbereitung.

### 3 Transfer von Fallstudien in Bildungsangebote

Der Transfer von Forschungsfallstudien in Bildungsangebote wird hier am Beispiel "Spanische Grippe" erläutert. Wir nutzen Methoden wie Fallstudien und problemorientiertes Lernen, um Datenkompetenzen zu vermitteln, die sowohl Wissen und Fähigkeiten als auch kritische Einstellungen einschließen. Die modulare Struktur und der Einsatz interaktiver Lehrbücher ermöglichen es den Lernenden Kontrolle über die Lernerfahrung zu geben und unterstützen dynamische Inhaltsaktualisierung.

#### 3.1 Von Bildungsbedarfen zu Fallstudien

Die digitale Transformation hat in den Geisteswissenschaften Konsequenzen für die Gegenstände, die Methoden, die Arbeitsweisen und die Infrastrukturen der Disziplinen. Vor allem digitale Methoden folgen einem anderen Paradigma, das auf der Formalisierung von Forschungsfragen und deren Operationalisierung für die computergestützte Analyse beruht [Mo13]. Dies führt zu einem dazu, dass bestehende Forschungsfragen neu aufgerollt werden, zum anderen wird durch die digital-quantitativen Methoden die Möglichkeit geschaffen, neue Forschungsfragen zu entwickeln, für die z.B. die Prozessierung einer großen Menge von Texten notwendig ist.

Angesichts dieser Entwicklung ist es entscheidend, dass Forschende nicht nur mit den notwendigen technischen Fähigkeiten ausgestattet werden, sondern auch eine kritische Perspektive auf die Möglichkeiten und Grenzen digitaler Methoden entwickeln. Die Vermittlung von Datenkompetenz umfasst nicht nur die Vermittlung von Wissen und Fähigkeiten, sondern auch die Förderung von Einstellungen und die Integration verschiedener Kompetenzen, die in die tatsächliche Forschungspraxis übertragbar sein sollten. Um solch komplexes Lernen [MKF24] zu erleichtern, setzen wir Methoden wie Fallstudien [Fo01] und problemorientiertes Lernen [Ka00] ein.

Fallstudien dienen dabei als Kern des Lehrkonzepts. Eine Fallstudie wird definiert als die Untersuchung einer konkreten Forschungsfrage oder eines spezifischen Problems innerhalb einer Studie. Sie beschreibt spezifische Aktivitäten oder Schritte aus der realen Forschungspraxis und bietet narrative Einblicke [Fo01]. Diese beinhalten oft detaillierte Hintergrundinformationen und zeigen den realen Prozess der Durchführung einer Studie, inklusive aller Herausforderungen und Unvollkommenheiten. Parallel dazu setzen wir auf problemorientiertes Lernen, bei dem die Lehrenden sich mit Aufgaben aus der Forschung auseinandersetzen [Ka00].

Im Folgenden wird exemplarisch auf eine Fallstudie eingegangen, die sich mit den Spuren der Spanische Grippe (1918-1920) in der zeitgenössischen Berliner Presse beschäftigt. Die konkret zu bearbeitende Problemstellung lautet dabei, ob sich die zeitliche Verlaufsform der medialen Thematisierung parallel zu den Grippewellen entwickelt. Ausgangspunkt für die Untersuchung bildet ein Zeitungskorpus aus dem Berliner Raum, das aus Scans histo-

rischer Zeitungen besteht, auf die über ZEFYS zugegriffen werden kann. Um die Zeitungen automatisch untersuchbar zu machen, muss das Korpus zunächst aufbereitet werden, indem der Text mittels OCR erkannt und anschließend unter Anwendung verschiedener Verfahren (manuell, regelbasiert, LLM-basiert) korrigiert wird. Anhand eines kleinen Auszugs wird daraufhin die Qualität des OCR-Outputs (Precision / Recall) bestimmt. Der Fokus der Fallstudie liegt auf der Aufbereitung und der Bereinigung des Untersuchungskorpus, wodurch die Wichtigkeit der Datenqualität für die Analyseergebnisse herausgestellt wird.

### 3.2 Lernziele und Fallstudien entlang des Datenlebenszyklus

Das Constructive Alignment-Modell [Bi96] wird als Lernenden-zentrierter Ansatz umgesetzt, der mit der Festlegung von Lernzielen beginnt und schrittweise zu weiteren Komponenten, Assessments und Lernaktivitäten übergeht. Die Lernziele werden in kognitive und affektive Dimensionen unterteilt. Kognitive Lernziele konzentrieren sich auf Denken, Wissen und Erkenntnisse, während affektive Lernziele Gefühle, Motivationen und Einstellungen betreffen, wie beispielsweise die Stärkung des ethischen Bewusstseins im Umgang mit Daten. Diese Kompetenzen werden in Grob- und Feinlernziele strukturiert. Groblernziele definieren übergeordnete Ziele für Fertigkeiten und Kenntnisse. Jedes Groblernziel wird in mehrere Feinlernziele unterteilt, die sich direkt aus den Anforderungen des Groblernziels ableiten [Ba14].

Feinlernziele werden präzise formuliert, um sie empirisch zu messen und effektive Assessments zu ermöglichen. Eine strukturierte Lernzielbeschreibung enthält einen Einleitungssatz, der den Kontext klärt, und spezifische Inhalte, die das Lernziel definieren. Bei deren Entwicklung kam die Bloom'sche Taxonomie [BI56] zum Einsatz, die eine klare, strukturierte Definition und eine angemessene Abfolge der Lernziele ermöglicht. Zusätzlich werden Bedingungen und Beurteilungskriterien wie Qualität, Menge oder Zeitmessung genau festgelegt, um die Zielerreichung zu bewerten [Bi19].

Die Entwicklung der Lernziele wurde durch die sequenzielle Struktur der empirischen Schritte/Aufgaben oder ausgewählten Aspekten der Fallstudie geleitet. Sie orientieren sich am Datenlebenszyklus und bilden die Struktur für die Bildungsangebote. Diese Identifizierung kleinerer Aufgaben innerhalb eines konkreten Problems basiert auf dem Workflow-basierten Lernen [MTS21] und dem Vier-Komponenten-Instructional-Design-Modell (4C/ID) [Me97]. Die Zentrierung des Lernprozesses um die Aufgaben fördert einerseits die aktive Beteiligung der Lernenden und spiegelt andererseits den sequenziellen Prozess der empirischen Forschungspraxis wider. Die Abfolge der Aufgaben entspricht den Inhalten und Strukturen der Groblernziele.

Tab. 1 veranschaulicht die Beschreibung der Grob- und Feinlernziele aus einem Ausschnitt für die Fallstudie "Spanische Grippe". Der Ausschnitt basiert auf Aufgabe/Schritt sechs aus 15 im Ablauf der Fallstudie, mit dem Fokus auf Anreicherung als einer Phase

des Datenlebenszyklus. Die Vorlage für die Beschreibung der Lernziele enthält zusätzlich die Bedingungen und Beurteilungskriterien für die Feinlernziele.

Beschreibung der Aufgabe in der Fallstudie	Groblernziel	Feinlernziel
Homogenisierung der Daten: Bild/PDF zu TXT: Mittels einer OCR-Pipeline werden aus den im Imageformat (JPG o.Ä.) oder als PDF vorliegenden Datei-en Textdaten erstellt und im TXT Format abgespeichert.	Nach Abschluss der Einheit sind die Lernenden in der Lage, Methoden der Homogenisierung von Textdaten unter Zuhilfenahme von Jupyter Notebooks anzuwenden und die Qualität der Datensammlung anhand eines Samples zu bewerten	<ol style="list-style-type: none"><li>1. Die Lernenden können die Grundprinzipien und den Workflow einer OCR-Pipeline erläutern. Sie können erklären, wie Bilder und PDF-Dokumente in bearbeitbare Textformate umgewandelt werden.</li><li>2. Die Lernenden können eine OCR-Pipeline in einem Jupyter Notebook ausführen, indem sie spezifische Python-Bibliotheken und Skripte verwenden, um Textdaten aus Bildern oder PDF-Dokumenten zu extrahieren.</li><li>3. Die Lernenden können die Schlüsselkonzepte und -terminologien, die zur Bewertung der Qualität einer OCR-Pipeline verwendet werden, klar und präzise erklären. Dazu gehören Konzepte wie Genauigkeit (Accuracy), Präzision (Precision), Wiederfindungsrate (Recall) und der F1-Score.</li><li>4. Die Lernenden können die Qualität der durch eine OCR-Pipeline erzeugten Daten anhand eines bereitgestellten Text-Samples quantitativ bewerten. Sie sind in der Lage, Python-Skripte in Jupyter Notebooks zu verwenden, um die Genauigkeit und andere relevante Metriken zu berechnen.</li><li>5. Die Lernenden können die Qualität der durch die OCR-Pipeline erzeugten Textdaten bewerten und fundierte Entscheidungen über die Angemessenheit der OCR-Ergebnisse für verschiedene Anwendungsfälle treffen.</li></ol>

Tab. 1: Grob- und Feinlernziele aus einem Ausschnitt für die Fallstudie “Spanische Grippe”

3.3 Mediendidaktisches Aufbereitung der Fallstudien

Die Struktur und die einzelnen Komponenten der Bildungsangebote orientieren sich am 4C/ID-Modell [Me97], das die Bedeutung von Aufgaben in den Mittelpunkt der Lernerfahrung stellt und gleichzeitig die Wichtigkeit von Interaktivität und Feedback betont. Da

den Lernenden keine direkte Anleitung durch Dozent:innen und keine direkt in die Lernmaterialien integrierte Möglichkeit zur Zusammenarbeit mit anderen geboten wird, fördert das Modell des Cognitive Apprenticeship [CBH91; WJ07] das Expert:innendenken in realen Forschungsszenarien durch eine explizite Darstellung. Die Materialien bereiten die Lernenden nicht nur darauf vor, die erarbeiteten Lösungen zu verstehen und anzuwenden, sondern decken auch die Komplexität der Entscheidungsfindung auf, beleuchten mögliche Fehlerquellen und bieten Einblicke in alternative Wege und Reflexionsmöglichkeiten.

Die Forschungsfragen oder deren Teile, die im Fokus stehen, und daraus resultierende Schritte oder Aufgaben einer Studie bilden das Zentrum der Lernerfahrung und strukturieren die Inhalte einer Fallstudie. Jedes Modul deckt eine spezifische Aufgabe und damit verbundenes Groblernziel ab, wobei die kleinsten Einheiten aus 1-3 Feinlernzielen bestehen, um eine effektive Modularisierung zu ermöglichen.

Module einer Fallstudie bestehen aus den folgenden Komponenten: Präsentation des Problems/der Aufgabe oder ihre Zusammenfassung (im letzten Modul), Unterstützende Informationen für Lösungsfindung, Interaktive Übungen, Diskussion und Reflexion, Just-in-time prozedurale Informationen, und Assessment. Während die Reihenfolge dieser Komponenten immer gleich ist, hängt ihre Kombination davon ab, in welchem Modul sie sich befinden: einleitendes Modul, inhaltsvermittelnde Module oder zusammenfassendes Modul.

**Präsentation der Aufgaben/Probleme und deren Zusammenfassung:** Jedes Modul einer Fallstudie beginnt mit der Darstellung der übergreifenden Forschungsfrage, des Kontexts, der Motivation zur Beantwortung der Forschungsfrage und, falls nötig, der Präsentation zusätzlicher Ressourcen (wie Datensätze, Datenerhebungsinstrumente usw.). Falls sich die Fallstudie nur auf bestimmte Schritte des Datenlebenszyklus konzentriert, wird dieser Fokus ebenfalls zu Beginn erklärt.

**Unterstützende Informationen (Theorie + Beispiele):** Nach der Erklärung der Aufgabe werden den Lernenden unterstützende Informationen zur Verfügung gestellt, die die Ausführung von Problemlösungs- und Denkaspekten dieser spezifischen Lernaufgaben fördern. Es wird beschrieben, wie der Aufgabenbereich organisiert ist und wie Probleme in diesem Bereich am besten angegangen werden können.

**Interaktive Übungen der Lernaufgaben/Probleme:** In inhaltsvermittelnden Modulen folgen darauf zwischen einer und vier interaktiven Übungen. Wenn diese kein unmittelbares automatisiertes Feedback erlauben, wird den Lernenden ausgearbeitete und erklärte Lösungen (als Text, Skript oder Video) bereitgestellt.

**Diskussion und Reflexion:** Damit die Lernenden die unterstützende Information eingehend verarbeiten und vertiefen können, wird durch Fragen zur Reflexion angeregt und eine weitere Diskussion gefördert. Jedes Modul schließt daher mit einer Reflexionskomponente ab, die Fragen stellt, Lösungen vergleicht sowie die Stärken und Grenzen der Ansätze erörtert und bei Abschlussmodulen auch nächste Schritte aufzeigt.

**Prozedurale Information (für Routineaspekte der Lernaufgaben):** Prozedurale Informationseinheiten erklären Lernenden, wie die Routineaspekte von Lernaufgaben auszuführen sind, und liefern How-to-Instruktionen (z. B. Anleitung zum Herunterladen oder Benutzen von Software). Da die Lernenden diese Informationen nicht auswendig lernen sollen, werden sie unmittelbar dort präsentiert, wo sie benötigt werden.

**Assessment:** Formatives Assessment wird innerhalb eines Moduls kontinuierlich durchgeführt, während summatives Assessment [Sc08] am Ende jedes Moduls stattfindet, um die Gesamtleistung zu evaluieren. Beide Bewertungsformen befinden sich noch in der Entwicklungsphase. Assessment dient im Rahmen der hier vorgestellten Lerneinheiten immer der Verbesserung der Lernerfahrung. Eine “Benotung” ist nicht vorgesehen, jedoch werden ggf. Hinweise für mögliche Prüfungsformen gegeben, sollten die Lerneinheiten für Zertifikatskurse o.ä. genutzt werden.

Als Format wurden interaktive Lehrbücher gewählt, die sich besonders für Forschende eignen. Dieses Format ermöglicht eine strukturierte und interaktive Präsentation von multimedialen Bildungsinhalten. Die hierarchische Struktur, bei der jedes Modul einem Kapitel und jede Sektion einem Kapitelabschnitt entspricht, ermöglicht Forschenden – sowohl in der Rolle als Lernende als auch als Lehrende, die diese Materialien für ihre Lehrtätigkeit nutzen – eine leichte Orientierung und Navigation. Eingebettete interaktive Übungen mit Datensätzen fördern praxisnahe Analysen und Problemlösungen, ohne dass eine umfangreiche Moderation oder Benutzerverwaltung erforderlich ist, was die Zugänglichkeit und das selbstgesteuerte Lernen unterstützt. Zudem erlaubt das Format den Lehrenden eine einfache Aktualisierung der Inhalte; neue Kapitel, Abschnitte oder Inhalte können leicht hinzugefügt oder durch externe Zusammenarbeit ergänzt werden.

Interaktive Lehrbücher können auf verschiedene Weise implementiert werden. Basierend auf internen Diskussionen und einem Vergleich unterschiedlicher Plattformen – Moodle, R-Shiny und Jupyter Books – haben wir uns aufgrund der Ergebnisse dieser Gegenüberstellung für Jupyter Books [CA22] entschieden (siehe Tab. 2).

Kriterien	Moodle	R + Shiny	Jupyter Books
Unterstützung interaktiver Elemente	Unterstützt effektiv interaktive Elemente wie Quizze. Erfordert jedoch Plugins für die Integration von Jupyter Notebooks, was zusätzlichen technischen Aufwand bedeutet.	Ermöglicht die direkte und effiziente Integration interaktiver Elemente, besonders geeignet für datenintensive Anwendungen.	Ermöglicht die Integration interaktiver Elemente durch Einbindung von Jupyter Notebooks. Unterstützt die Ausführung und Anpassung von Quizzen, Code, Gleichungen und Visualisierungen.
Feedback-mechanismen	Bietet umfangreiche Optionen für sofortiges Feedback durch	Primär für datenintensive, interaktive	Ermöglicht Feedback durch Integration von Jupyter Notebooks



Kriterien	Moodle	R + Shiny	Jupyter Books
	Quiz- und Aufgabenmodule.	Visualisierungen entwickelt. Benutzerdefinierte Feedbacksysteme können mit entsprechendem Programmieraufwand implementiert werden.	mit Modulen wie Jupyter-quiz oder Plugins.
Flexibilität der Ausführung	Kann über Plugins mit externen Tools verzahnt werden, was jedoch zusätzlichen technischen Aufwand erfordert.	Die Konfiguration erfordert zusätzliches technisches Know-how und ist nicht immer direkt über die Benutzeroberfläche zugänglich.	Bietet Flexibilität in der Ausführungsumgebung und ermöglicht den Wechsel zwischen Cloud-basierten Umgebungen, lokaler Datenverarbeitung und Live-Code-Ausführung.
Kompatibilität	Bietet grundlegende Integration mit verschiedenen Tools und Programmiersprachen.	Ist in R integriert und kompatibel mit R-basierten Analyse-Tools und Programmierumgebungen.	Ermöglicht die direkte Einbindung von Jupyter-Notebooks, die Programmiersprachen wie Julia, Python und R unterstützen. Für zusätzliche Sprachen sind Plugins erforderlich.
Benutzerfreundlichkeit für Lehrende	Weit verbreitet und bietet viele Inhaltsoptionen für einfache Einbindung durch webbasierte Navigation. Breite Nutzerbasis und ausgereifte Dokumentation.	Erfordert R-Kenntnisse für die Einrichtung. Nutzer:innen mit R-Erfahrung können interaktive Elemente integrieren. Basiskenntnisse in Shiny-Syntax sind für Benutzeroberflächengestaltung notwendig.	Setzt technische Kenntnisse für die Administration voraus. Nutzer:innen, die Jupyter Notebooks kennen, integrieren diese problemlos. Grundlegende Markdown-Kenntnisse für einfache Text-Inhalte sind erforderlich.

Tab. 2: Vergleich von Moodle, R-Shiny und Jupyter Books: Auswahlkriterien basierend auf internen Diskussionen und didaktischen Ansätzen

Auf der Startseite des Jupyter Books für die Fallstudie “Spanische Grippe”<sup>7</sup> wird die Liste der Module angezeigt, die auch im Menü (links) permanent zugreifbar ist. Die Fallstudie wird entweder sequentiell durchlaufen, oder einzelne Module werden direkt angesteuert. Im Sinne des Aktiven Lernens [Ma16] kann der Lernpfad auf diese Weise selbst konstruiert werden. Die Suche bietet zusätzlich die Möglichkeit, gezielt bestimmte Inhalte anzu-steuern. Zudem kann auf das GitHub-Repository des Jupyter Books zugegriffen werden, das Book kann lokal gespeichert werden, oder einzelne Sektionen können ausführbar ge-macht werden.

Die Übersichtsseite sowie das Modul “Corpus Collection” sind einleitende Module zur Fragestellung und zum Untersuchungskorpus sowie dazugehörigen Metadaten. Die fol-genden Module sind inhaltsvermittelnd. In ihnen wird zuerst textuell in ein Problem ein-geführt, dann werden unterstützende Informationen bereitgestellt (Abb. 2), woraufhin in-teraktive Übungen folgen. Abb. 3 zeigt eine Übung, in der manuell korrigierter Text ein-geben und mit dem durch OCR automatisch erzeugten Text verglichen wird. Damit wird an Performance-Metriken herangeführt. Am Ende des Moduls bzw. der Fallstudie werden die Inhalte zusammenfassend diskutiert und reflektiert, z. B. im Hinblick auf die verwen-deten Daten, Methoden und Tools (Abb. 4).

**How OCR works**

**Die Grippe wütel weiter**  
Zunahme der schweren Fälle in Berlin.

Die Zahl der Grippefälle ist in den letzten beiden Tagen auch in Groß-Berlin noch erheblich gestiegen. Die Warenhäuser und son-  
stigen großen Geschäfte, die Kriegs- und die pri-  
vaten Betriebe klagen, daß übermäßig viele An-  
gestellte sich haben krank melden müssen, und auch  
bei der Post und bei der Straßenbahn ist der  
Prozentatz der Grippekranken bedentend ge-  
stiegen.

A modern Optical Character Recognition (OCR) algorithm typically involves several stages. Here's a breakdown of the key stages:

- 1. Preprocessing:** This initial step involves preparing the image for analysis and recognition. Common preprocessing tasks include:
  - Noise Reduction:** Removing noise from the image to enhance the text's clarity. This could involve filtering techniques like Gaussian blur or median filter.
  - Binarization:** Converting the image from grayscale or color to black-and-white, where text is typically represented as black pixels on a white background. This helps in distinguishing the text from the background.
- 2. Images into digital text.** OCR


[What OCR tools are there](#)

- 2.1. Evaluate OCR engine quality
- 2.2 Process the whole corpus of PDF-s with the same OCR engine
- 2.3. OCR postprocessing
3. Getting digital text from the structured markup (XML)

Now let's use all the data for processing and analysis (next notebook)

Abb. 2: Unterstützende Informationen (GIFs und Text) im Modul *Corpus Preparation*

<sup>7</sup> Die aktuellste Version ist zugänglich unter: <https://dh-network.github.io/quadrige/markdown/intro.html>



Waves of the Spanish Flu – Case Study

Corpus Collection

Introduction: Texts as digital objects

Metadata – Conceptual Introduction

Corpus Preparation

Introduction to Optical Character Recognition (OCR)

**Data Input and Homogenisation**

OCR quality

### 2.1.1 Manually create the 'ground truth' to evaluate against

```
ground_truth = input('Please insert corrected string: ')
```

run restart restart & run all

Please insert corrected string:

```
print(ground_truth)
```


run restart restart & run all

Die Lage auf dem Kohlenmarkte gibt zu den schlimmsten Befürchtungen Anlass. Für Sachsen fehlten im November 30.000 Wagen zu je 10 Tonnen und für Dezember wird mit noch größeren Ausfällen gerechnet werden. Es ist mit einem völligen Stillstand der Industrie innerhalb vierzehn Tagen zu rechnen, wenn nicht eine erhebliche Steigerung der Belastungen der Kohlenbergwerke oder ihrer Zahl gelingt. Weiter steht eine wesentliche Erhöhung der Kohlenpreise bevor.

Contents

1. Types of input data for text corpora
- 2. Images into digital text. OCR**
  - How OCR works
  - What OCR tools are there
  - 2.1. Evaluate OCR engine quality**
    - 2.1.1 Manually create the 'ground truth' to evaluate against**
    - 2.1.2 Measure OCR precision, recall and F-measure
    - Precision in OCR
    - Recall in OCR
    - F-measure (F1 Score) in OCR
  - 2.2 Process the whole corpus of PDF-s with the same OCR engine
  - 2.3. OCR postprocessing
  3. Getting digital text from the structured markup (XML)

Abb. 3: Interaktive Übung als Texteingabe im Modul *Corpus Preparation*



Waves of the Spanish Flu – Case Study

Corpus Collection

Introduction: Texts as digital objects

Metadata – Conceptual Introduction

Corpus Preparation

Introduction to Optical Character Recognition (OCR)

Data Input and Homogenisation

OCR quality

**Post-Correction of the OCR Output**

Introduction to the Post-Correction of OCR output

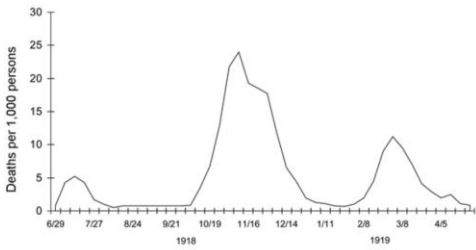
Rule-based Correction OCR output

LLM-based correction of OCR output

### Discussion of the intermediate result

Is this result meaningful and does it actually reflect something? One way to check that is to compare our plot with the actual data about the intensity of the pandemic.

In (Taubenberger, J. K., & Morens, D. M. (2006). 1918 Influenza: the Mother of All Pandemics. *Emerging Infectious Diseases*, 12(1), 15-22. <https://doi.org/10.3201/eid1201.050979>) it is stated that 'The first pandemic influenza wave appeared in the spring of 1918, followed in rapid succession by much more fatal second and third waves in the fall and winter of 1918–1919, respectively'. They also supplement this statement with a plot from an earlier paper (Jordan E. (1927). *Epidemic influenza: a survey*. Chicago: American Medical Association):



Our two waves of mentions of the word 'Grippe' seem to correspond to the mortality figures, which could indicate that the method, albeit very simple, works and that historical events can sometimes be reflected in word frequency counts...

Abb. 4: Zusammenfassende Diskussion der Resultate im Modul *Corpus Analysis*

3.4 Bereitstellung als offene Bildungsangebote

Die auf diese Weise aufbereiteten Fallstudien sollen für zwei verschiedene Zielgruppen bereitgestellt werden, was zwei korrespondierende Distributionswege mit sich bringt:

- Für Lehrende sind eine leistungsfähige Suche nach bestimmten Themen und/oder Formaten, ein einfacher Download in die eigene Sammlung und eine flexible Integration in die jeweils vorhandenen Lehr-/Lernumgebungen (bspw. institutionelle Lernplattformen) nötig. Präferierte Zielplattformen für die Veröffentlichung der Bildungsangebote sind daher OER-Repositoryen.
- Für Lernende ist ebenfalls eine leistungsfähige Suche erforderlich. Der anschließende Zugriff sollte aber keinen Download erforderlich machen, sondern auch direkt auf der Plattform möglich sein. Damit werden ggf. auch zusätzliche Funktionen wie Lernfortschrittskontrollen oder Kollaboration möglich.

Für die Fallstudie “Spanische Grippe” ist angesichts der Flexibilität der Jupyter Books beides über dasselbe Werkzeug möglich. Für andere Fallstudien können die Anforderungen zur Werkzeugauswahl jedoch andere Umsetzungen sinnvoller erscheinen lassen – etwa als interaktiver Moodle-Kurs, der sowohl instantiiert in einer offenen Moodle-Plattform (für Lernende) als auch exportiert als übertragbarer Kurs in einem Repository (für Lehrende) bereitzustellen wäre. Für beide Distributionswege stellt jedoch die Metadatenbasierte Suche gleichermaßen den Einstiegspunkt dar.

## 4 Zusammenfassung und Ausblick

Die Vermittlung von Datenkompetenzen erfordert sowohl adäquate Bildungsangebote als auch dauerhaft tragfähige Strukturen, die sich an den Bedarfen und Möglichkeiten der Forschung orientieren. Der hier vorgestellte Ansatz bietet dreierlei Beiträge zu dieser Herausforderung. Die Verzahnung von Forschungs-, Lern- und Vernetzungsorten entlang von Fallstudien liefert erstens eine strukturelle und prozessuale Basis für die Bereitstellung von Bildungsangeboten. Zweitens erlaubt das präsentierte Vorgehen zur didaktischen Aufbereitung von forschungsbasierten Fallstudien in mediengestützte Bildungsangebote eine gleichermaßen effiziente wie qualitätsorientierte Produktion von Lehr-/Lernmaterial. Die Implementierung als interaktive Lehrbücher wurde exemplarisch als Jupyter Book “Spanische Grippe” erläutert, ist jedoch unmittelbar auf andere Themen oder Werkzeuge übertragbar. Drittens liefern wir mit den Überlegungen zur Integration in eine verteilte Struktur von Repositoryen und Lernumgebungen einen Beitrag zur Weiterentwicklung des digitalen Bildungsraums [LKW23].

Der Fokus der weiteren Arbeiten liegt auf der Vervollständigung und Erprobung aller beschriebenen Komponenten, einschließlich Assessment und Feedback. Die Modularisierung und Verknüpfung verschiedener Fallstudien bzw. darauf basierender Bildungsangebote stellt die nächste Herausforderung dar. Schließlich wird die Integration in eine wohldefinierte Architektur – i.S.e. plattformübergreifenden Integration in ein zusammenhängendes Ganzes [Kn22] – über die dauerhafte Tragfähigkeit der Angebote entscheiden.

## Danksagung

Die hier vorgestellten Arbeiten wurden teilweise gefördert vom Bundesministerium für Bildung und Forschung in den Projekten FoLD und QUADRIGA unter Förderkennzeichen 16DHB3018 bzw. 16DKZ2034A/16DKZ2034H.

## Literaturverzeichnis

- [Ba14] Baumgartner, P.: Taxonomie von Unterrichtsmethoden: ein Plädoyer für didaktische Vielfalt. Waxmann, Münster New York München Berlin, 2014.
- [Bi96] Biggs, J.: Enhancing teaching through constructive alignment. *Higher Education* 3/32, S. 347–364, 1996.
- [Bi19] Bilon, E.: Using Bloom's Taxonomy to Write Effective Learning Objectives. The ABCDs of Writing Learning Objectives: A Basic Guide. Independently published, Independently published, 2019.
- [Bl56] Bloom, B.S. et al. Hrsg.: Taxonomy of educational objectives. The classification of educational goals: Handbook I: Cognitive Domain. David McKay Company, New York, 1956.
- [Bu24] Buchholz, B. et al.: Umsetzungskonzept QUADRIGA: Berlin-Brandenburgisches Datenkompetenzzentrum für Digital Humanities, Verwaltungswissenschaft, Informatik und Informationswissenschaft. Zenodo, 2024.
- [CA22] Chen, E.; Asta, M.: Using Jupyter Tools to Design an Interactive Textbook to Guide Undergraduate Research in Materials Informatics. *Journal of Chemical Education* 10/99, S. 3601–3606, 2022.
- [CBH91] Collins, A.; Brown, J. S.; Holum, A.: Cognitive Apprenticeship: Making Thinking Visible. *American Educator: The Professional Journal of the American Federation of Teachers* 15, 1991.
- [Cr16] Crymble, A. et al.: The Programming Historian - Print Edition. Zenodo, 2016.
- [Fo01] Foran, J.: The Case Method and the Interactive Classroom. *Thought & Action* 1/17, S. 41–50, 2001.
- [HBK18] Heidrich, J.; Bauer, P.; Krupka, D.: Future Skills: Ansätze zur Vermittlung von Data Literacy in der Hochschulbildung. *Hochschulforum Digitalisierung*, Arbeitspapier 37, 2018.
- [Ka00] Kay, J. et al.: Problem-Based Learning for Foundation Computer Science Courses. *Computer Science Education* 2/10, S. 109–128, 2000.
- [Kn22] Knoth, A. et al.: Structural Challenges in the Educational System meet a Federated IT-Infrastructure for Education – Insights into a Real Lab. In: *Proc. 14th Int. Conf. on Computer Supported Education (CSEDU)*, 369-375, 2022.

- [LKW23] Lucke, U.; Knoth, A.; Wilhelm-Weidner, A.: Perspektiven von Wissenschaft und Praxis auf die digitale Vernetzungsinfrastruktur für die Bildung. *e-learning and education*, 15/1, 2023.
- [Lu24] Lucke, U. et al.: Digitalisierungsbezogene Kompetenzen in den Fächern entwickeln - Fachintegriert, interdisziplinär oder fachunabhängig? In: *Forschen | Lernen - Digital*. Universitätsverlag Potsdam. Im Erscheinen, voraussichtlich 2024.
- [Ma16] Markant, D. B. et al.: Enhanced Memory as a Common Effect of Active Learning. *Mind, Brain, and Education* 3/10, S. 142–152, 2016.
- [Me97] van Merriënboer, J. J. G.: Training complex cognitive skills. A four-component instructional design model for technical training. *Educational Technology Publ*, Englewood Cliffs, N.J, 1997.
- [MKF24] van Merriënboer, J. J. G.; Kirschner, P. A.; Frèrejean, J.: Ten steps for complex learning. A systematic approach to four-component instructional design. Routledge, New York, NY, 2024.
- [Mo13] Moretti, F.: 'Operationalizing': or, the function of measurement in modern literary theory. *Pamphlets of the Stanford Literary Lab*, 6, 2013.
- [MTS21] Mischke, D.; Trilcke, P.; Sluyter-Gäthje, H.: Workflow-basiertes Lernen in den Geisteswissenschaften: digitale Kompetenzen forschungsnah vermitteln: Bildung in der digitalen Transformation. Waxmann, Münster, New York, S. 190–195, 2021.
- [Ri15] Ridsdale, C. et al.: Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report. [https://www.researchgate.net/publication/284029915\\_Strategies\\_and\\_Best\\_Practices\\_for\\_Data\\_Literacy\\_Education\\_Knowledge\\_Synthesis\\_Report](https://www.researchgate.net/publication/284029915_Strategies_and_Best_Practices_for_Data_Literacy_Education_Knowledge_Synthesis_Report), Stand: 16.07.2023.
- [SBH19] Schüller, K.; Busch, P.; Hindinger, C.: Future Skills: Ein Framework für Data Literacy. Zenodo, 2019.
- [Sc08] Schermutzki, M.: Lernergebnisse - Begriffe, Zusammenhänge, Umsetzung und Erfolgsermittlung. Lernergebnisse und Kompetenzvermittlung als elementare Orientierungen des Bologna-Prozesses. In (Benz, W.; Kohler, J.; Landfried, K. Hrsg.): *Handbuch Qualität in Studium und Lehre: Evaluation nutzen, sichern, Profil schärfen*. Raabe, Berlin, S. 1–30, 2008.
- [SS24] SSH Open Marketplace | SSHOPENCLOUD. <https://sshopencloud.eu/ssh-open-marketplace>, Stand: 14.05.2024.
- [WJ07] Woolley, N. N.; Jarvis, Y.: Situated cognition and cognitive apprenticeship: a model for teaching and learning clinical skills in a technologically rich and authentic learning environment. *Nurse education today* 1/27, S. 73–79, 2007.