

# *Distant reading – Tools and methods*

December 12-13, 2019 - Basel University



Français	1
Inscription	1
Objectif	1
Programme	2
Deutsch	3
Anmeldung	3
Inhalt	3
Programm	4
Italiano	5
Registrazione	5
Contenuti	5
Programma	6
Conveners	7

## Français

### Inscription

Lieu : Universitätsbibliothek Basel, Vortragssaal (1. Stock) ; Schönbeinstrasse 18-20, CH-4056 Basel

Date : 12 et 13 décembre 2019

Les doctorants des universités de Bâle, Berne, Fribourg, Genève, Neuchâtel, et Lausanne (UNIL et EPFL) sont prioritaires, mais les postdoctorants et les étudiants (motivés) de master sont les bienvenus. La participation est gratuite sur inscription.

Pour les doctorants, les frais de transports sont remboursés, ainsi que la nuit d'hôtel pour les participants domiciliés à plus d'une heure de Bâle.

Langue d'enseignement : anglais.

### Objectif

La lecture distante (*distant reading*), portée par le développement du numérique dans les sciences humaines, s'est imposée comme une des approches les plus prolifique des textes littéraires. Les cartes, les graphiques et les arbres, pour reprendre les mots de Moretti (2005), nous permettent en effet de relire les œuvres les plus célèbres d'une manière inédite, ou de nous pencher sur des textes jusqu'alors oubliés. De nouveaux motifs apparaissent, des hypothèses peuvent être confirmées pour la première fois à l'aide d'études systématiques sur de grands corpus. Ces études nouvelles ne reviennent cependant que trop peu sur l'importance de la collecte des données : d'où viennent-elles ? Comment sont-elles construites ?

La présente école se propose de revenir sur l'étape cruciale de l'acquisition des données, en revenant dans le détail sur leur chaîne de production. Nous commencerons par l'OCR (*optical character recognition*, reconnaissance optique de caractère), qui permet de transformer un jeu d'image en un texte exploitable, en dépit des difficultés introduites par la variation des systèmes graphiques ou la matérialité des artefacts anciens. Le second temps – décisif – est celui de l'encodage en XML-TEI, qui transforme le texte en base de données exploitable et permet d'ajouter un surcroît d'information au texte (auteur, genre, période). Le troisième et dernier temps est celui de l'analyse avec R, qui permet de tester des hypothèses par l'analyse et la visualisation de données.

Fortement tournée vers la pratique, cette école voudrait jeter les bases d'un premier corpus suisse multilingue (français, italien et allemand). Le cours sera l'occasion de discuter de sa fabrication, mais aussi des enjeux de la lecture distante.

Ce cours s'inscrit dans le cadre de l'action européenne COST *Distant Reading for History Literary History* [<https://www.distant-reading.net>]

## Programme

- Jour 1
  - 9h30-10h : Accueil des participants
  - 10h-10h30 : Introduction à la lecture distante
  - 10h30-11h : Introduction à l'OCR
  - 11h-11h30 : Pause-café
  - 11h30-12h30 : TP : entraîner et utiliser un OCR
  - 12h30-14h : repas
  - 14h-15h30 : Introduction à la construction de corpus : balancement du corpus + TP
  - 15h30-16h : Pause-café
  - 16h-17h30 : Introduction à la construction de corpus : encoder son corpus + TP
  - 18h-19h : Keynote I, *The Challenges of Composing a Swiss Literature Corpus* (Rosemarie Zeller, UNIBAS)
- Jour 2
  - 9h-10h : Keynote II, *The Challenges of Encoding Multi-lingual Corpora* (Carolin Odebrecht, Berlin HU)
  - 10h-10h30 : Pause-café
  - 11h-12h30 : Encodage de corpus (avancé) + TP
  - 12h30-14h : repas
  - 14h-16h : Encodage de corpus + TP
  - 16h-16h30 : Pause-café
  - 16h30-17h30 : analyse des résultats

## **Deutsch**

### **Anmeldung**

Standort : Universitätsbibliothek Basel, Vortragssaal (1. Stock) ; Schönbeinstrasse 18-20, CH-4056 Basel

Datum : 12. und 13. Dezember 2019

Doktoranden der Universitäten Basel, Bern, Freiburg, Genf, Neuenburg und Lausanne sowie der EPFL können sich für diesen Kurs anmelden. Post-Doc-Forscher können sich bewerben.

Die Teilnahme an dieser Veranstaltung ist für Doktoranden kostenlos. Die Reise- und Aufenthaltskosten werden durch das Doktorandenprogramm übernommen.

### **Inhalt**

Distant Reading, ein Verfahren, das durch die Digitalisierung in den Geisteswissenschaften entstanden ist, hat sich als einer der produktivsten Ansätze für literarische Texte erwiesen. Karten, Grafiken und Bäume, so Moretti (2005) in seinem Buch „Graphs, Maps, Trees: Abstract Models for a Literary History“ ermöglichen die innovative Relektüre berühmter Werke ebenso die Beschäftigung mit in Vergessenheit geratenen Texten. Neue Muster werden sichtbar, Hypothesen können erstmals systematisch auf grösseren Korpora überprüft werden. Jedoch wird beim Distant Reading oftmals die wichtige Ebene der ursprünglichen Datenerfassung vernachlässigt: Woher kommen die Daten? Wie werden sie gewonnen? Welche Implikationen haben hier bestimmte Entscheidungen?

Unser Kurs schlägt vor, zur entscheidenden Phase der Datenerfassung zurückzukehren, indem wir die Produktionskette detailliert beschreiben. Wir beginnen mit OCR (Optical Character Recognition), ein Verfahren, das einen Bilddatensatz in nutzbaren Text umwandelt, wobei Variationen in Druck, Orthographie sowie Materialität der Artefakte Herausforderungen darstellen. Die zweite – und entscheidende – Einheit ist die XML-TEI-Codierung, die die gewonnenen Textdaten in eine durchsuchbare Datenbank transformiert und mit weiteren Informationen, etwa zu AutorIn, Gattung und Publikationszeitraum, versieht. Als dritte Einheit wird die Analyse mit der Software R aufgezeigt, die es ermöglicht, Forschungsfragen zu testen, sowie Daten explorativ zu analysieren und zu visualisieren.

Stark praxisorientiert möchte dieser Kurs den Grundstein für ein erstes mehrsprachiges Schweizer Literaturkorpus (Französisch, Italienisch und Deutsch) legen. Anhand dieses Korpus wird es im Verlauf des Kurses Gelegenheit geben, das Verfahren des Distant Reading und seine Bedingungen auf allen Ebenen zu diskutieren.

Der Kurs ist angebunden an ein gemeinsames Forschungsprojekt im Rahmen des europäischen Projekts Distant Reading for European Literary History <https://www.distant-reading.net/>.

## Programm

- Tag 1
  - 9h30-10h : Empfang
  - 10h-10h30 : Einführung „Distant Reading“
  - 10h30-11h : Einführung in OCR
  - 11h-11h30 : Kaffeepause
  - 11h-12h30 : TP: Einführung und Verwendung OCR
  - 12h30-14h : Mittagessen
  - 14h-15h30 : Einführung in den Korpusbau: Repräsentativität und Balancing + TP
  - 15h30-16h : Kaffeepause
  - 16h-17h30 : Einführung in den Korpusbau: XML-Kodierung + TP
  - 18h-19h : Keynote I, *The Challenges of Composing a Swiss Literature Corpus* (Rosemarie Zeller, UNIBAS)
- Jour 2
  - 9h-10h : Keynote II, *The Challenges of Encoding Multi-lingual Corpora* (Carolin Odebrecht, Berlin HU)
  - 10h-10h30 : Kaffeepause
  - 11h-12h30 : XML-Kodierung (erweitert) + TP
  - 12h30-14h : Mittagessen
  - 14h-16h : XML-Kodierung + TP
  - 16h-16h30 : Kaffeepause
  - 16h30-17h30 : Analyse der Ergebnisse

## Italiano

### Registrazione

Luogo: Universitätsbibliothek Basel, Vortragssaal (1. Stock) ; Schönbeinstrasse 18-20, CH-4056 Basel

Date: 12 e 13 Dicembre 2019

I dottorandi delle Università di Basilea, Berna, Friburgo, Ginevra, Neuchâtel e Losanna (UNIL ed EPFL) avranno la priorità, ma i ricercatori post-doc e gli studenti delle lauree magistrali sono i benvenuti. La partecipazione è gratuita.

Per gli studenti di dottorato (se domiciliati a più di un'ora di viaggio da Basilea) vengono rimborsate le spese di trasporto e una notte in hotel.

Lingua di insegnamento: inglese

### Contenuti

La lettura distante (*distant reading*), guidata dallo sviluppo delle tecnologie digitali nelle scienze umane, si è imposta come uno degli approcci più prolifici per lo studio dei testi letterari. Mappe, grafici e alberi, per usare le parole di Moretti nel suo libro *Graphs, Maps, Trees: Abstract Models for a Literary History*, ci permettono di rileggere le opere più famose in un modo nuovo, o di riscoprire testi precedentemente dimenticati. Nuovi schemi emergono, e ipotesi possono essere confermate per la prima volta con l'aiuto di studi sistematici su grandi corpora. Questi nuovi studi, tuttavia, si soffermano troppo poco sull'importanza della raccolta dei dati: da dove vengono? Come sono costruiti?

Questa scuola mira a tornare alla fase cruciale dell'acquisizione dei dati, soffermandosi in dettaglio sulla loro catena di produzione. Inizieremo con il riconoscimento ottico dei caratteri (OCR), che trasforma un'immagine in un testo utilizzabile, nonostante le difficoltà introdotte dalla variazione dei sistemi grafici o dalla materialità dei manufatti antichi. La seconda fase - decisiva - è quella della codifica XML-TEI, che trasforma il testo in un database ricercabile e rende possibile aggiungere ulteriori informazioni al testo (come autore, genere e periodo). La terza e ultima fase è quella dell'analisi con R, che consente di verificare le ipotesi tramite l'analisi e la visualizzazione dei dati.

Fortemente orientata alla pratica, la scuola ambisce a gettare le basi per un primo corpus multilingue svizzero (francese, italiano e tedesco). Il corso sarà un'occasione per discutere della sua fabbricazione, ma anche dei problemi della lettura distante.

Il corso è collegato a un progetto di ricerca congiunto realizzato nell'ambito della COST Action europea *Distant Reading for History Literary History* [<https://www.distant-reading.net>].

## Programma

- Giorno 1
  - 9:30-10:00 Accoglienza dei partecipanti
  - 10:00-10:30 Introduzione al distant reading
  - 10:30-11:00 Introduzione all'OCR
  - 11h-11h30 Pausa caffè
  - 11:30-12:30 Esercitazione: allenare e utilizzare un OCR
  - 12:30-14:00 Pranzo
  - 14:00-15:30 Introduzione alla costruzione del corpus: bilanciamento del corpus + esercitazione
  - 15:30-16 Pausa caffè
  - 16:00-17:30 Introduzione alla costruzione del corpus: codifica del corpus + esercitazione
  - 18:00-19:00 Keynote I, *The Challenges of Composing a Swiss Literature Corpus* (Rosemarie Zeller, UNIBAS)
- Giorno 2
  - 9:00-10:00 Keynote II, *The Challenges of Encoding Multi-lingual Corpora* (Carolin Odebrecht, Berlin HU)
  - 10:00-10:30 Pausa caffè
  - 10:30-12:30 Codifica del corpus (avanzata) + esercitazione
  - 12:30-14:00 Pranzo
  - 14:00-16:00 Codifica del corpus + esercitazione
  - 16:00-16:30 Pausa caffè
  - 16:30-17:30 Analisi dei risultati

## Conveners

**Elias Kreyenbühl** (PhD) is leading the digitization at the Basel University Library. Since 2009 Elias is actively engaged in the digitization of cultural heritage. Today he is enabling the scholarly use of digitized historical documents. As a historian he cares about critical thinking and digital literacy.

**Simone Rebor**a holds a PhD in Foreign Literatures and Literary Studies (University of Verona) and a BSc in Electronic Engineering (Polytechnic University of Torino). Currently, he works as a research fellow between the University of Verona and the DH Lab of the University of Basel. His main research interests are theory and history of literary historiography and reader response studies. In the field of digital humanities, he focused on tools and methods like OCR, stylometry, and sentiment analysis.

**Dr. Berenike Herrmann** is PI of the Basel SNF-Digital Lives Project Forschungslogiken in den texbasierten DH nach dem Machine Learning Turn as well as Management Committee member of the COST Action Distant Reading for European Literary History and the current chair of the ADHO Special Interest Group “Digital Literary Stylistics”. As a postdoctoral researcher (“Oberassistentin”) at the DHLab at the University of Basel (Switzerland), her research focuses on the digital modeling of literary discourse. Topics are valuation of literature by users of social reading platforms, the style of writers of the 19th and early 20th Centuries, as well as the representation of “nation”, “cosmopolitanism”, and “region” in the German language continuum.

**Simon Gabay** holds a PhD in latin philology (universiteit van Amsterdam) and is currently employed as a post doctoral researcher at the université de Neuchâtel, where he teaches digital humanities. His main research interests are digital philology and the history of manuscripts. He is particularly interested in using NLP tools for literary studies and data extraction.