An early career research group focusing on the study of historical text reuse. Text reuse describes the spoken and written repetition of content.

**RESEARCH QUESTIONS & OBJECTIVE**

1. How does the human mind identify a reuse unit in context? Can machines do the same?
2. How can we detect text reuse across languages (big data)?
3. How reliable are online and offline retrieval systems in delivering stable result-sets when looking for reuse at scale?

**Goal:** The investigation of stability and volatility of text reuse and its primitives. **Objective:** To create a multilingual dataset in order to train text reuse algorithms to detect reuse units across languages and at scale.

**TRACER**

eTRAP’s TRACER is a suite of 700 algorithms, whose features can be combined to create the optimal formula for detecting those words, sentences and ideas that have been reused across texts. Specifically, the algorithms look for **primitives**, the stable elements of reuse units.

TRACER’s framework comprises **six steps**.

**CURRENT DATASET: FAIRY TALES**

1. **Snow White** (AT 709): Grimm [DE], Pushkin [RU], Tsvetaeva [RU], Calvino [IT], Jacobs [EN], Bruford [EN], Campbell [EN], Taylor [EN], Briggs [EN].
2. **Puss in Boots** (AT 545B): Grimm [DE], Straparola [IT], Pushkin [RU].
3. **The Fisherman and his Wife** (AT 555): Grimm [DE], Wild [DE], Pommern [DE], Pushkin [RU], Briggs [EN], Keding [EN], Andreev [UA].

**TEXT REUSE AT SCALE**

We conducted experiments with idioms as text reuse units. We used two approaches to identify these idioms at scale:

A. Google Custom Search (online): search for full and condensed version of idioms in Google Books and the web.

B. Apache Lucene (offline): search for full and condensed version of idioms in zeno.org (DARIAH-DE), Deutsches Textarchiv and Project Gutenberg.

**Observations:** The warmer the colour, the more similar the results. Lucene delivers significantly better results than Google when searching for full and condensed versions of idioms in text collections. This is because Google compromises on accuracy to favour speed, and because Lucene allows the user to set retrieval parameters for higher accuracy.