

Characterizing Knowledge Manipulation in a Russian Wikipedia Fork [a.k.a. RWFork]

First study of how original Wikipedia content has been forked and manipulated to meet the requirements of a national regulation.



Universitat
Pompeu Fabra
Barcelona



Authors:

Mykola Trokhymovych
Oleksandr Kosovan
Nathan Forrester
Pablo Aragon
Diego Saez-Trumper
Ricardo Baeza-Yates

Introduction

Wikipedia runs on MediaWiki, a free and open-source platform that also powers many other wiki-based encyclopedias. One of them is the recently launched Ruwiki (RWFork), which copied and modified Russian Wikipedia content to comply with Russian law. This article offers a comprehensive comparative analysis of meta-information and geographical, temporal, categorical, and textual features to examine changes made by RWFork editors.

Main Findings

- Only 1.75% of articles are targeted, but they attract 14.2% of views, showing a focus on high-interest, often controversial topics.
- Editing peaks during weekday office hours and drops on weekends, indicating coordinated, possibly professional activity.
- Around 90% of edits fall into 8 categories; 44% relate to the Russo-Ukrainian War. Other topics include culture, gender, and medical laws.

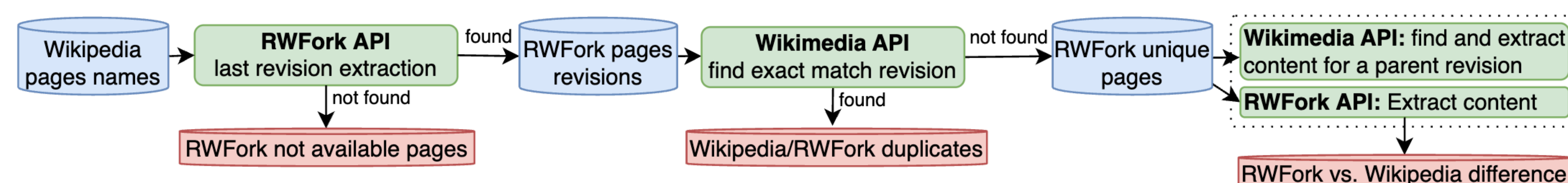
Challenges

- Extracting fine-grained content differences between forked and original articles.
- Collecting and aligning additional metadata (e.g., geography, engagement) from multiple sources;
- Building a meaningful taxonomy of changes using LLMs without labeled data or supervision.

Contributions

- A novel methodology to compare an alternative Wikipedia version with its source.
- An open-source dataset of all content modifications introduced by RWFork, enabling further research.
- A comprehensive analysis of changes, along with the taxonomy of knowledge manipulation practices.

Data collection & preparation

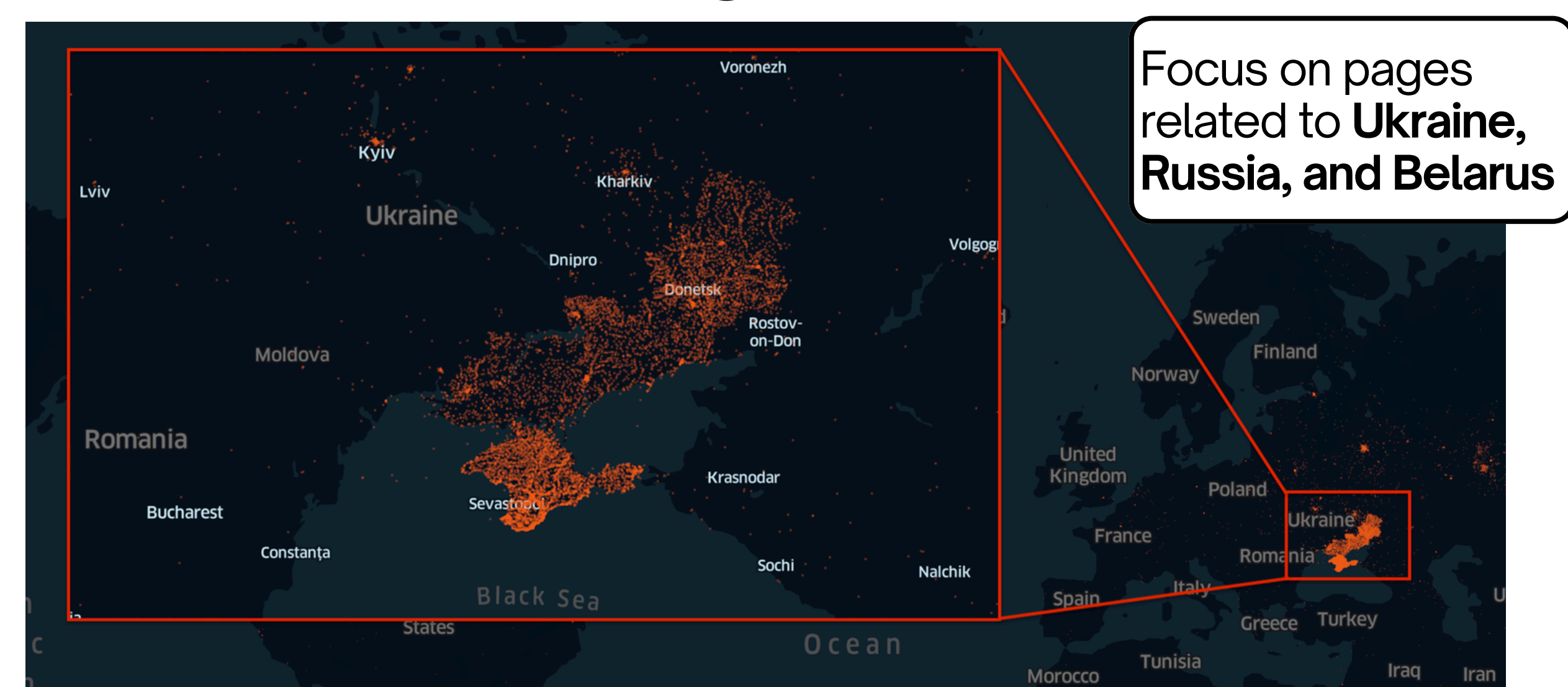


RWFork Dataset characteristics:

- Size:** 33,664 changed articles
- Language:** Russian
- Time:** May and September 2023

How? Matching RWFork articles with their source and extracting the diff.

Location of changed articles



Pages statistics comparison



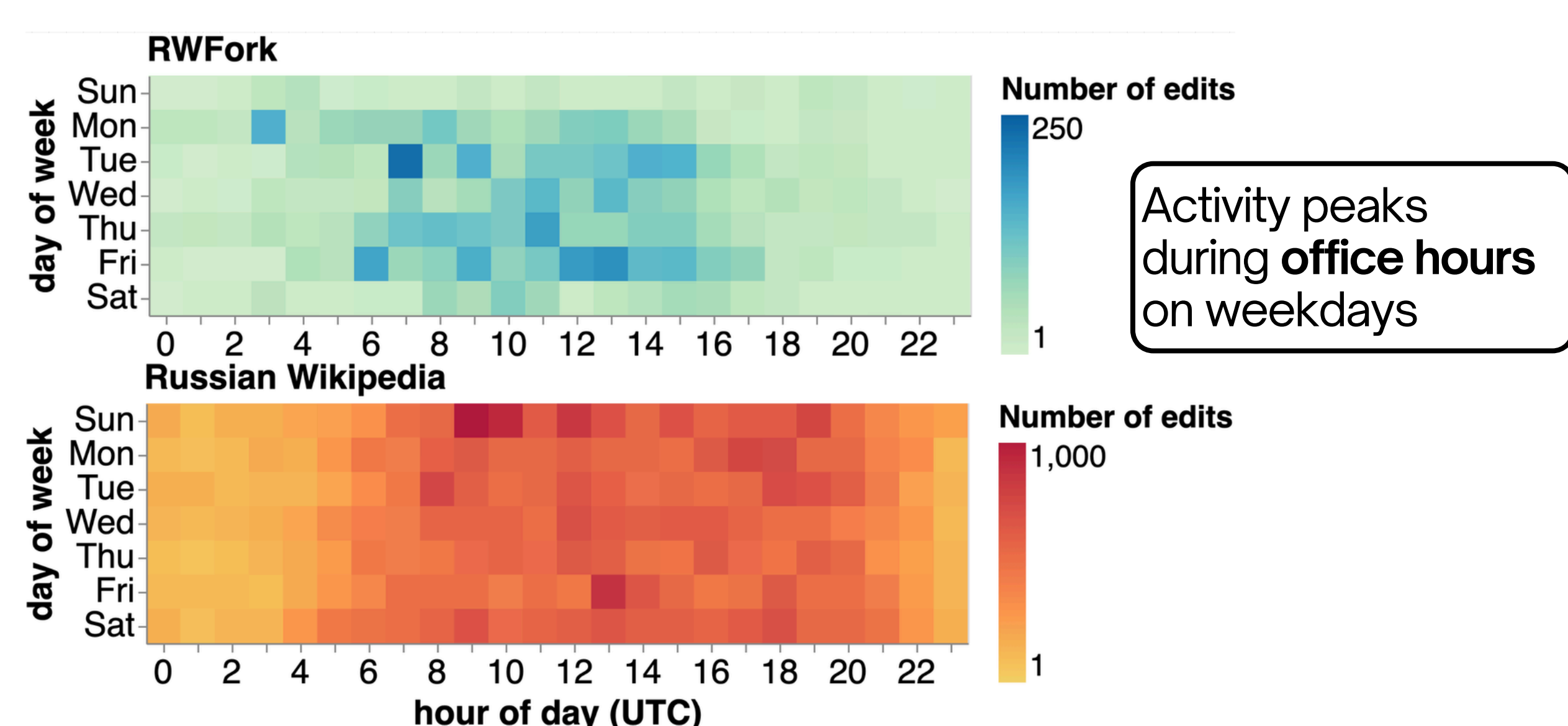
Group:

- changed
- duplicated
- missing

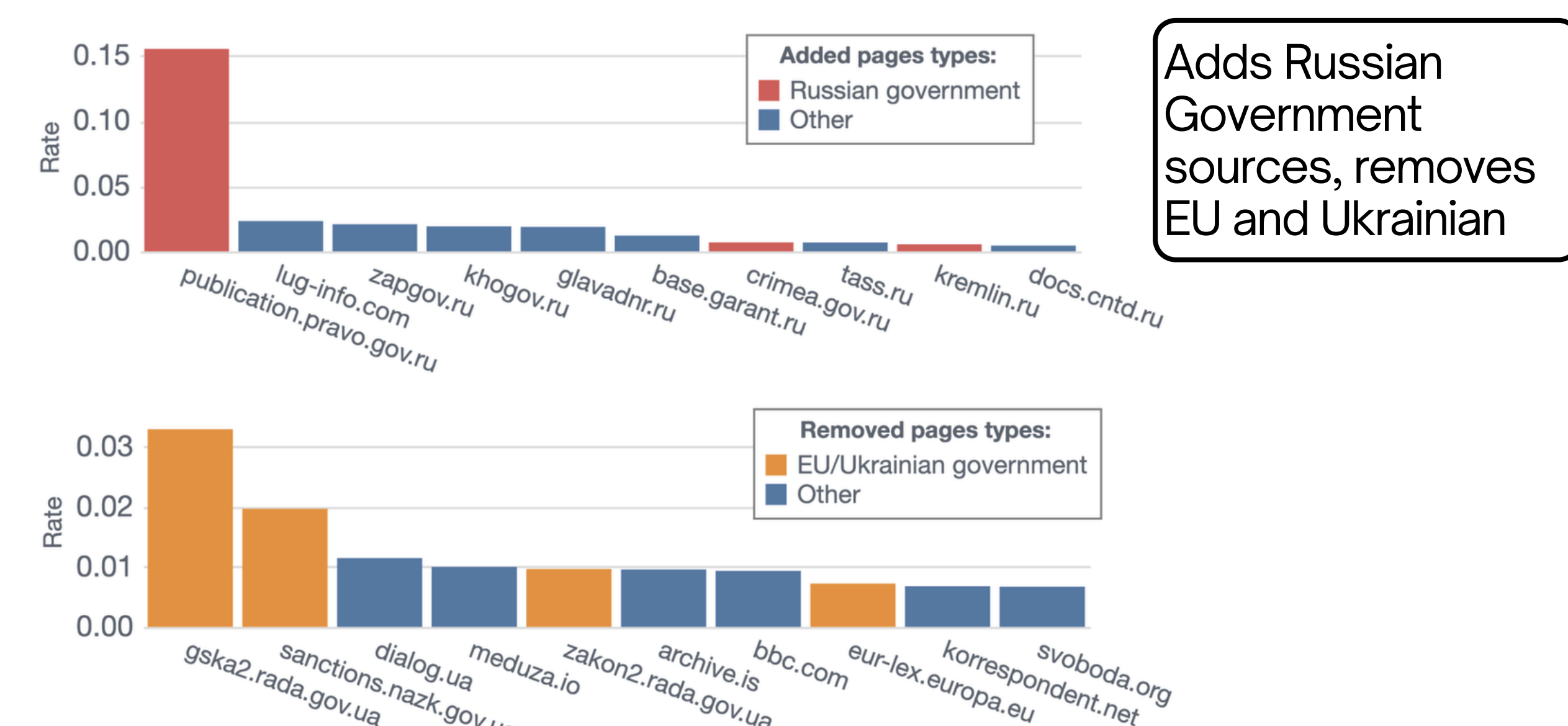
Changing influential and controversial pages

14.2% page views
~x2 revert rate
~x5 edits on avg.

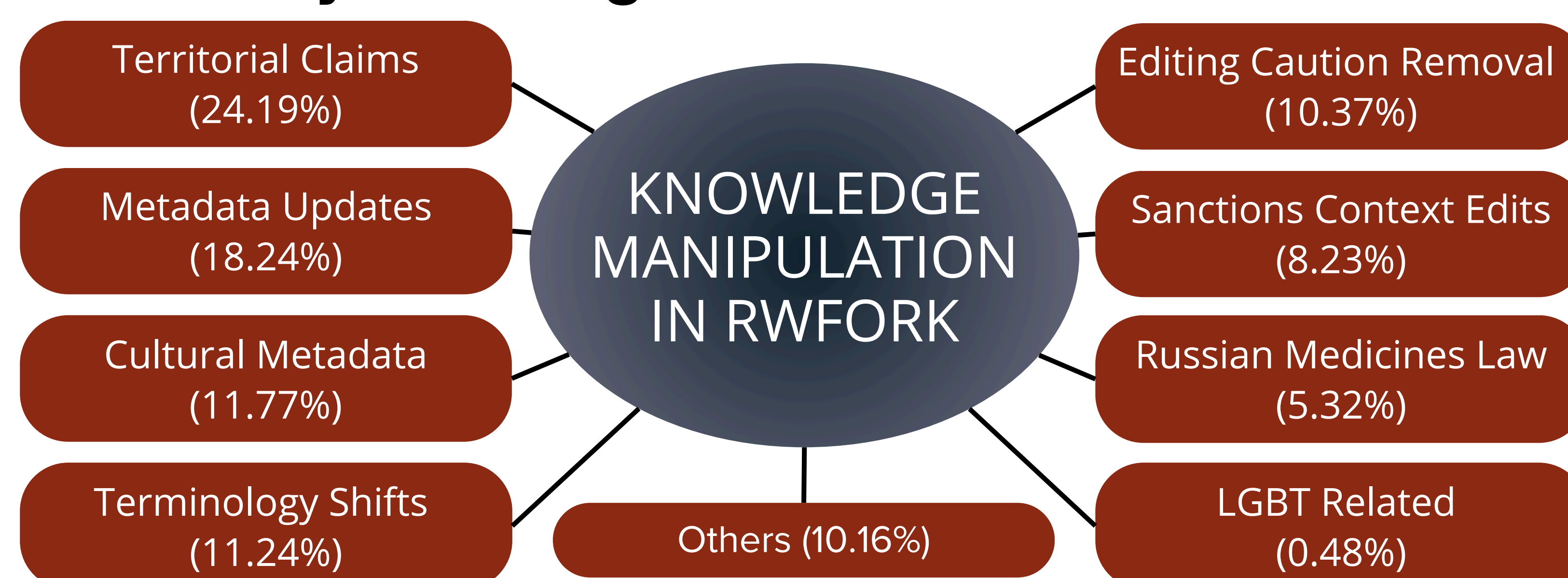
Temporal Editing Patterns



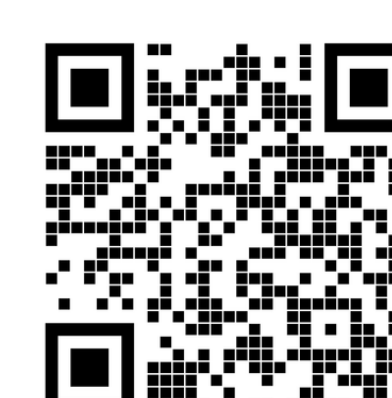
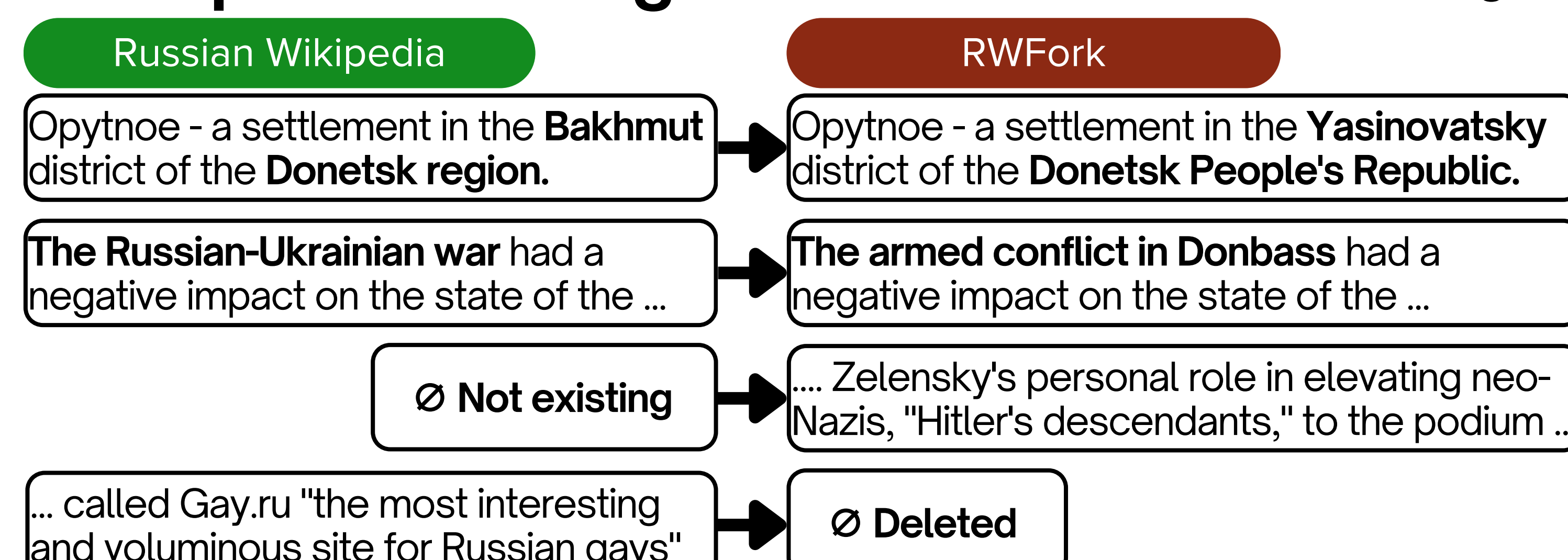
References Editing Patterns



Taxonomy of changes



Examples of changes [translated from Russian to English]



Dataset