iLCM – A Virtual Research Infrastructure for Large-Scale Qualitative Data

Andreas Niekler[1], Arnim Bleier[2], Christian Kahmann[1], Lisa Posch[2,3], Gregor Wiedemann[1], Kenan Erdogan[2], Gerhard Heyer[1], Markus Strohmaier[4,2]

[1]Faculty of Mathematics and Computer Science,University Leipzig,
Augustusplatz 10, 04109 Leipzig, Germany
{aniekler, kahmann, wiedemann, heyer}@informatik.uni-leipzig.de

[2]Department of Computational Social Science,
GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8 50667 Cologne, Germany
{firstname.lastname}@gesis.org

[3]Institute of Interactive Systems and Data Science,
Graz University of Technology,
Inffeldgasse 16c, 8010 Graz, Austria

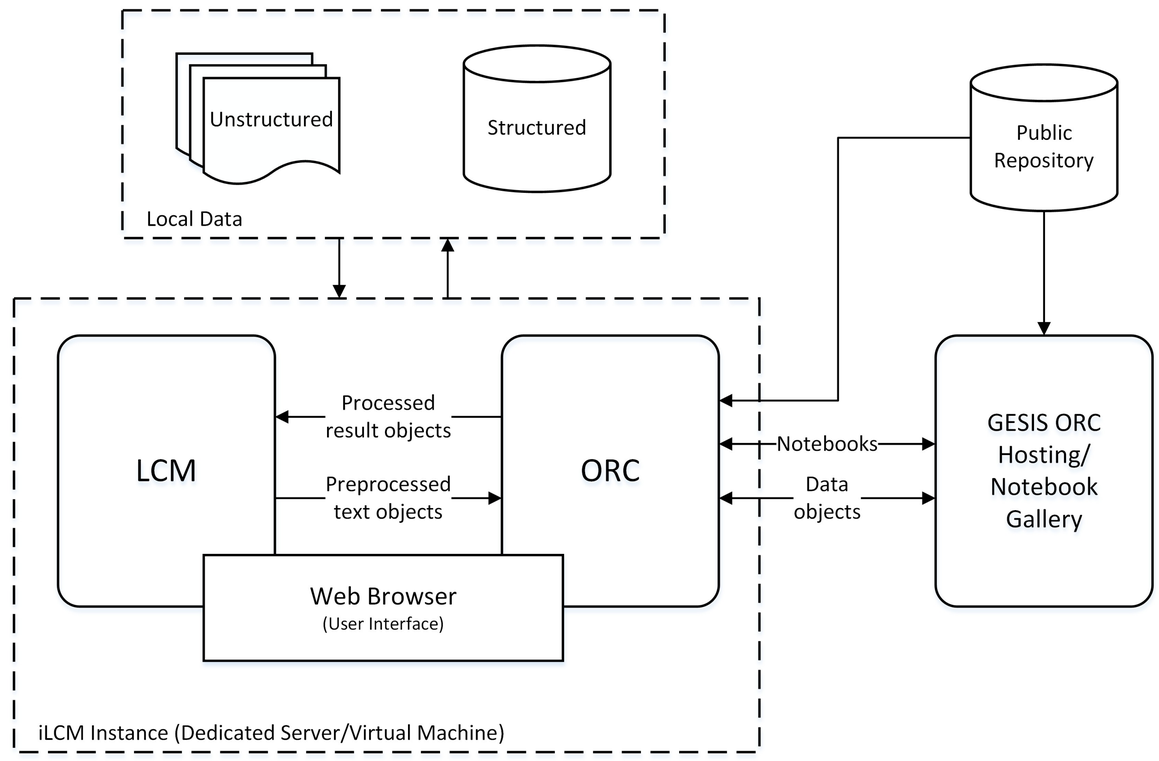
[4]HumTec Institute, RWTH Aachen University
Theaterplatz 14, 52062 Aachen, Germany
markus.strohmaier@humtec.rwth-aachen.de

Computational social science (CSS) is the interdisciplinary study of socio-cultural phenomena through new kinds of data and technologies. One of its central objectives is the **extraction of useful and interpretable knowledge from potentially large behavioral digital datasets**.

In CSS, **unstructured data (usually text)** and **structured data (e.g. metadata, survey data or log/sensor data)** are both important sources of information. Through the digitization of social processes, they are available in large quantities. With the iLCM infrastructure, **we want to offer analysis capabilities for reusable and reproducible research with both data types**.

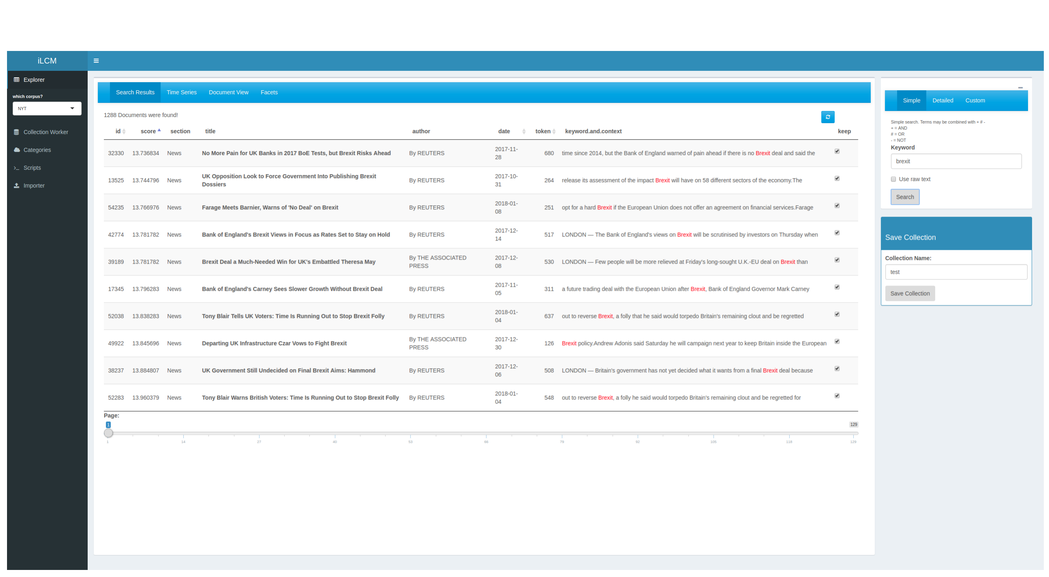
Users of computer-assisted text analysis from these disciplines face non-trivial technical and methodological challenges, especially when it comes to the requirement to analyze large amounts of text and reproducibility of research. The iLCM infrastructure provides solutions to these challenges.

The iLCM architecture consists of two major components, the **Leipzig Corpus Miner (LCM) to preprocess and analyze large amounts of unstructured text data, and the ORC component to further process structured data extracted from text or imported from external sources.** Analysis scripts on structured data, so-called **notebooks, will be published in a central repository**. The LCM provides a browser-based **Graphical User Interface (GUI) targeting end-users from applied disciplines who are primarily interested in the application of generic text mining methods on their data**. The **Open Research Computing (ORC) environment runs scripts in active executable documents, so-called "notebooks".** By this, the ORC component allows analyzing structured data either extracted from texts by the LCM or imported from external database.

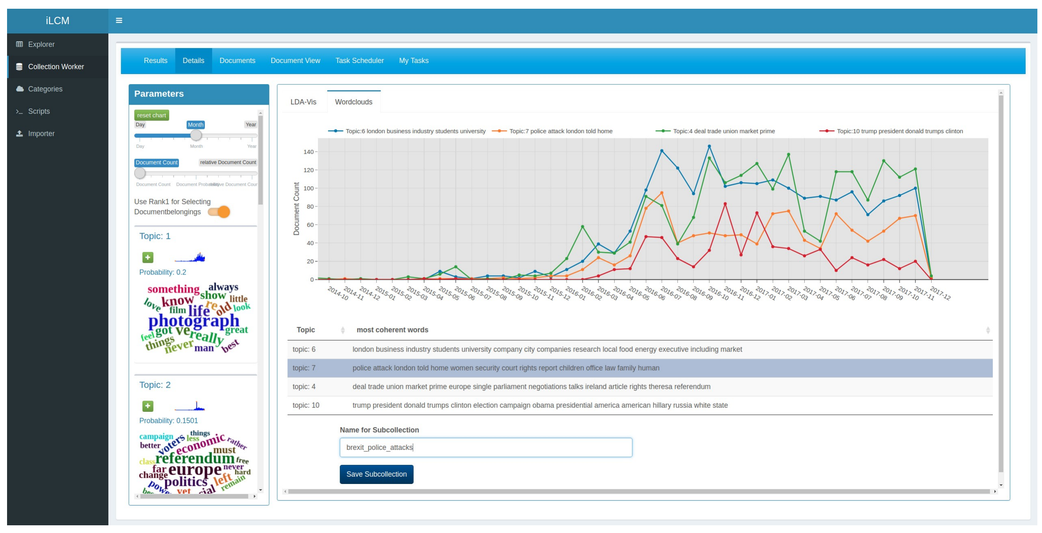


Leipzig Corpus Miner (LCM)

The LCM is not a stand-alone program, but rather a **server infrastructure** comprising a number of components including a **document database** (MariaDB), an **NLP pipeline** for preprocessing text data (spaCy), **a full-text index** (Solr), a **collection of text mining processes** and finally a **web application GUI** (R Shiny).



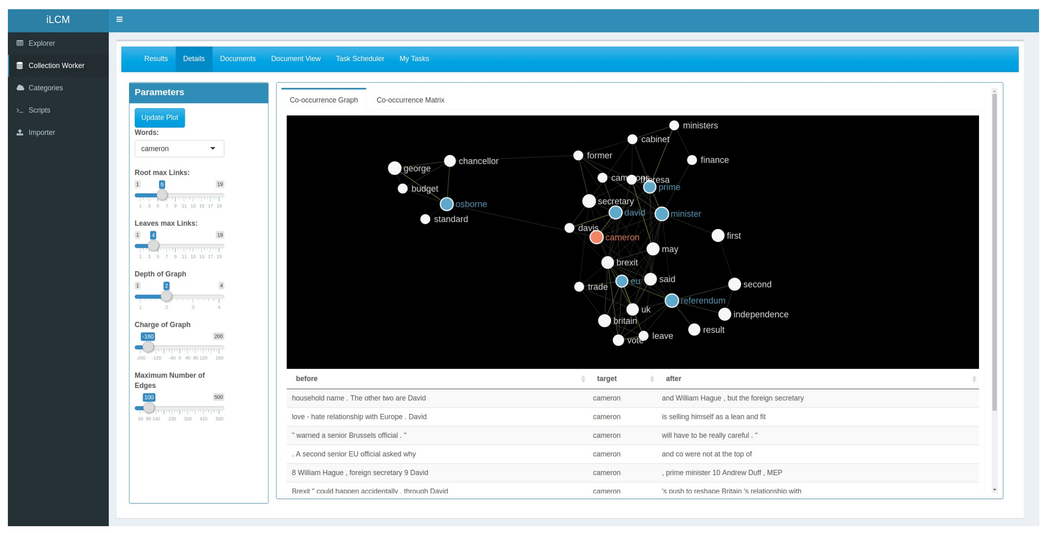
LCM Document Search. Screenshot of the document search interface.



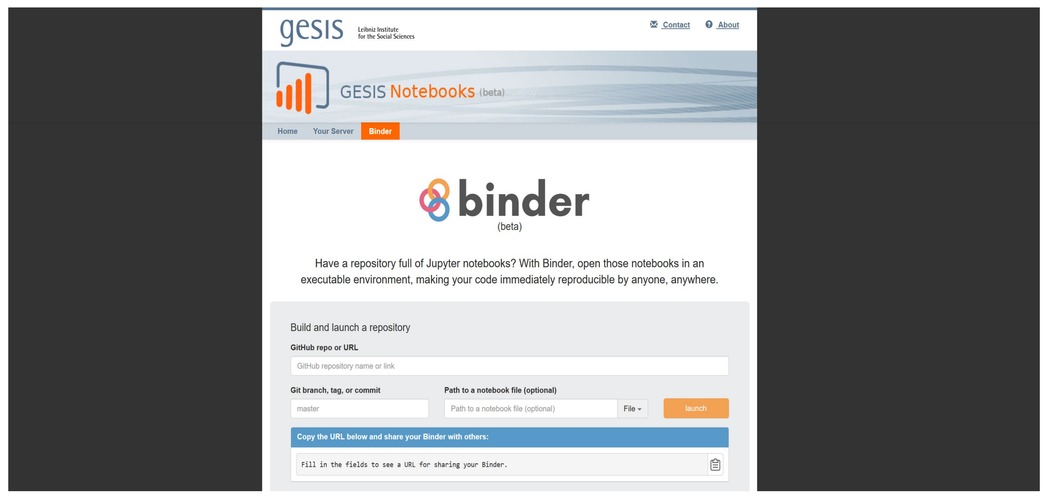
LCM Topic Model Results. Screenshot of a topic model visualization.

Open Research Computing (ORC)

The ORC component extends the iLCM with an **editor environment for program scripts** which enables the **creation of scripts along with their documentation**. GESIS will host a public instance of the ORC environment. This ORC instance is then **extended with a repository on which notebooks can be published** by users. The repository not only **secures the long-term availability** of published notebooks, it also serves as a notebook gallery that **users can browse and search via keywords and metadata**.



LCM Word Context / Co-occurrence Results. Screenshot of word context and co-occurrence graph visualizations.



Instance of an ORC-environment hosted by GESIS. It is based on Jupyter Notebooks and binder (https://binderhub.readthedocs.io/en/latest/).

Project information

The project iLCM is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in the funding line for virtual research environments (FKZ/project number: 324867496).
Funding period: 02/2017–01/2020

Research institutions:

GESIS – Leibniz Institute for the Social Sciences
Department of Computational Social Science,
Dr. Arnim Bleier

University Leipzig
Faculty of Mathematics and Computer Science
Department of Natural Language Processing
Prof. Dr. Gerhard Heyer

ilcm.informatik.uni-leipzig.de