Tiepmar, Jochen (2018): Implementation and Evaluation of the Canonical Text Service Protocol as Part of a Research Infrastructure in the Digital Humanities.
Von der Universität Leipzig angenommene Dissertation zur Erlangung des akademischen Grades Doctor Rerum Naturalium im Fachgebiet Informatik.

Zusammenfassung der Dissertation

Einer der bestimmenden Faktoren moderner Gesellschaften ist die fortlaufende Digitalisierung von Informationen und Ressourcen. Dieser Trend spiegelt sich in heutiger Forschung wider und hat starken Einfluss auf akademische und industrielle Projekte. Es ist nahezu unmöglich, ein modernes Projekt aufzusetzen, welches keinerlei digitale Aspekte beinhaltet und viele Projekte werden mit dem alleinigen Zweck der Digitalisierung eines Teils der Welt ins Leben gerufen. Dieser Trend führt zur Entstehung neuer Forschungsfelder an den Schnittstellen zwischen der analogen Welt – beispielsweise den Geisteswissenschaften – und der Digitalen – beispielsweise der Informatik. Eine davon ist das für diese Arbeit interessante Gebiet der Digital Humanities.

Dabei werden komplexe Forschungsfragen, -techniken und -prinzipien verbunden, die sich unabhängig voneinander entwickelten. Viel Mühe ist nötig, um die Kommunikation zwischen deren Konzepte zu definieren um Missverständnisse und Fehleinschätzungen zu vermeiden. Dieser Prozess der Brückenbildung ist eine zentrale Aufgabe der neu entstehenden Forschungsfelder.

Diese Arbeit schlägt eine solche Brücke für die textorientierten Digital Humanities vor. Diese Lösung basiert auf einem Referenzsystem für digitalen Text, welches in den Geisteswissenschaften spezifiziert und im Rahmen dieser Arbeit zu einem Datenkommunikationsprotokoll für die Informatik uminterpretiert wurde: dem Canonical Text Service (CTS) Protokoll.

Während dieser Arbeit wird das Protokoll mit Hinblick auf geisteswissenschaftliche/editorische und technische Aspekte analysiert und es wird herausgearbeitet, welche Vorteile sich beide Parteien von ihm versprechen können. Es wird eine hocheffiziente Implementierung beschrieben und anhand eines wiederverwendbaren Performanzbenchmarks evaluiert.

Zusätzlich wird CTS mit Hinblick auf zwei Trends analysiert, welche aus der fortlaufenden Digitalisierung folgen: Big Data und Interoperabilität.

Wissenschaftlicher Beitrag

- Die Haupterrungenschaft dieser Arbeit ist eine hocheffiziente Implementierung eines webbasierten feingranularen Textreferenzsystems, wie es durch die geisteswissenschaftliche Community gewünscht wurde. Diese wurde explizit darauf hin entwickelt, ohne außergewöhnliches technisches Know-How verwendbar zu sein.
- Das CTS Protokoll wird aus einer technischen und einer geisteswissenschaftlichen Perspektive heraus analysiert, was beiden Parteien hilfreiche Einblicke in die jeweils anderen Ansichten geben kann. Dabei wird CTS auch in den Kontext von zwei aktuellen technischer Trends gesetzt: Big Data und Interoperabilität.
- Es wird eine Analyse der aktuellen und absehbaren Big Data Anforderungen der textorientierten Digital Humanities durchgeführt.
- Eine datenzentrische Analyse des CTS Protokolls liefert Grundlagen für weitere Implementierungen. Praktische Implementierungsskizzen mit verschiedenen Datenmodellen zeigen CTS-spezifische Fallen und Probleme auf.
- Die Verknüpfung der CTS-Implementierung mit der Forschungsinfrastruktur CLARIN liefert Forschern einen unkomplizierten Weg, ihre Daten über ein etabliertes System einem breiten Publikum zugänglich zu machen und dabei eine Technik zu nutzen, die sie möglicherweise sowieso nutzen wollten.
- Eine Reihe von Text Mining Tools und Anwendungen wurden im Kontext dieser Arbeit entwickelt, inklusive einem skalierbaren Workflow zur Zitationsanalyse, strukturbasiertes Textalignment und einem frei zugänglichen peristent ziterbaren Text Mining Framework.
- Die in dieser Arbeit durchgeführte Performanzevaluierung liefert einen ersten vergleichbaren und implementierungsunabhängig gestalteten Performanztest. Dieser kann dazu genutzt werden, zukünftige CTS Implementierungen oder Einflüsse von Optimierungen miteinander zu vergleichen.

Tiepmar, Jochen (2018): Implementation and Evaluation of the Canonical Text Service Protocol as Part of a Research Infrastructure in the Digital Humanities.
Von der Universität Leipzig angenommene Dissertation zur Erlangung des akademischen Grades Doctor Rerum Naturalium im Fachgebiet Informatik.

Summary of the PHD-Thesis

One of the defining factors of modern societies is the ongoing digitization of information, resources and in many ways even life itself. This trend is obviously also reflected in today's research environments and heavily influences the direction in which academic and industrial projects are headed. It is borderline impossible to set up a modern project without including digital aspects and many projects are even set up for the sole purpose of digitizing a specific part of the world. One of the side effects of this trend is the emergence of new research fields at the intersection points between the analog world – represented for example by the humanities – and the digital world – represented for example by computer science. One set of such research fields are the digital humanities, the area of interest for this work.

In the process of this development, complex research questions, techniques, and principles are aligned next to each other that were developed independently from another. A lot of work has to go into defining communication between the concepts to prevent misunderstandings and misconceptions on both sides. This bridge building process is one of the major tasks that must be done by the newly developed research fields.

This work proposes such a bridge for the text-oriented digital humanities based on a digital text reference system that was previously developed in the humanities and is in this work reinterpreted as a data communication protocol for computer science: The Canonical Text Service (CTS) protocol.

In this thesis, the protocol is analyzed based on humanistic/editorial and technical requirements and it is discussed, what benefits both parties can expect from it. A highly efficient implementation is discussed and evaluated using a reusable implementation-independent performance benchmark.

In addition, CTS is analyzed with respect to two trends that are a result of the vast number of digitization projects: Big Data and Interoperability.

Scientific Contributions

- The main achievement of this work is a highly efficient implementation of an online fine-grained text reference system, as requested by the humanistic research community, which was designed with the explicit requirement of being accessible without advanced technical knowledge. This means that it is relevant to and instrumental for both computer scientists and (digital) humanists.

- The CTS protocol is analyzed from both a computer scientific and a humanistic perspective. This may provide useful insights into the other parties' philosophies and requirements. This analysis also puts into context two major trends in computer science: Big Data and Interoperability.

- It provides a Big Data analysis of current and foreseeable developments in the text-oriented digital humanities.

- A data-centric analysis of the CTS protocol provides the groundwork for future implementations and sketches out implementations in other data models or formats, describing potential pitfalls.

- The CTS connection to the CLARIN infrastructure provides a comparatively uncomplicated way for digital humanists to make their data available to a broad research community, based on digitization work that they may have intended to do anyway.

- Certain text mining tools and applications were developed during this project. This includes structure-based real-time text alignment, a citation analysis workflow that can deal with large documents, and a publicly available, persistently citable text mining framework.

- The performance evaluation of this implementation provides a baseline with a standardized data set and a comprehensive set of test scenarios that are developed specifically to test various performance-relevant circumstances. Since it is independent of the proposed CTS implementation, it can be used to compare future systems.