Scenario-based planning for the semantic digitization of historical reference works

Werner Scheltjens[1] (0000-0002-5209-9052), Christoph Schlieder[2] (0000-0002-7226-8204)
[1]Digital History, [2]Cultural informatics
University of Bamberg, Germany

Abstract

The retrodigitization of library collections that are exempt from copyright restrictions has made a myriad printed reference works instantly available in digital form, but the use of these collections is limited by the characteristics of the 'first digitization', which typically comprises a scan, an automatically produced OCR text and metadata. Developing the case of a famous nineteenth-century metrological reference work, this contribution shows the potential benefits of a second, or semantic, retrodigitization for historical research. The 'second digitization' aims to extract and model the semantic structure of encyclopaedic knowledge. Semantic modelling of digitized historical reference works is planned with the aid of a scenario-based methodology, which builds on and extends the planning approach suggested by Lodi et al (2017). The paper develops and employs competency questions to decide where to represent semantic relations explicitly and where to infer them with queries.

Extended abstract

The retrodigitization of library collections that are exempt from copyright restrictions has made a myriad printed reference works instantly available in digital form, encouraging historians with various specializations to familiarize with novel methods for using these sources (Paju et.al. 2020). Historical metrologists, for example, share a particular interest in eighteenth- and nineteenth-century reference works about industry and trade that tried to capture systematically the knowledge of their time. Such works were widespread, especially during the transition to the metric system (until ca. 1870). They combined a rather positivistic interest in measures, weights and currencies across the world with an attempt to meet the requirements of increasing standardization and systematization of societies (Kramper 2019). Like all historical texts, metrological reference works and lexica are witnesses of a particular historical period. In digital form, they constitute novel sources for historical research.

A famous example is Christian and Friedrich Noback's renowned and much-cited *Vollständiges Taschenbuch der Münz-, Maass- und Gewichtsverhältnisse* (Noback & Noback 1850; Denzel 2002; Witthöft 2018). The Bavarian State Library has digitized this reference work. Scans in pdf-format and the automatically OCR'ed contents of the book are available for non-commercial use. The 'first retrodigitization' has produced a digital version of the *Vollständiges Taschenbuch* that significantly simplifies the reading of single entries in the book. Novel insights about the composition of historical metrological systems, however, can only be obtained through comparative analysis of a large number of articles in the book, for example, the grain measures in all articles about places belonging to the same economic region. Scrolling through the scans or searching in the OCR'ed text are inefficient, if not impossible, methods for conducting this kind of analysis. Even though historical reference works, such as the *Vollständiges Taschenbuch*, are now available in digital form, they still await the semantic disclosure of their contents.

We argue that a second, or semantic, retrodigitization is necessary for answering specific research questions in the realm of historical metrology. The second retrodigitization complements the first, largely automatic retrodigitization and aims at the explicit modelling of the semantic structure of encyclopaedic knowledge. In principle, we can rely on the general methods of ontological modelling to plan the second digitization and make semantic relations in our reference work explicit. Most importantly, we refer to scenario-based methods that relate the planning process to so-called compentency questions, i.e. questions that specialist users of the source would like to examine and answer with the help of a data model (Kendall, McGuinness, 2019). Lodi et al. (2017) and Carriero et al. (2021) have successfully adapted this renowned scenario-based method to questions of the Digital Humanities. They solved the issue of modelling the metadata of Italian heritage institutions by recurring to competency questions for ontology development.

We show that this approach has its limitations when applied to research questions of digital history. A first issue results from the difference between modelling approaches that address metadata and such that rely on (primary) source data. Based on the planning of the semantic modelling of the Nobacks' *Vollständiges Taschenbuch*, we show that competency questions are necessary to decide where to represent semantic relations explicitly and where to infer them with queries. Moreover, our research makes clear that domain ontologies that have not originated in the context of DH, e.g. the metrological ontologies for the natural sciences (Martín-Recuerda et al. 2020), can hardly be used directly. We propose a scenario-based process that builds on and extends several aspects of the planning approach suggested by Lodi et al (2017).

Based on an example from the realm of historical metrology, a workflow for explicating semantic relations in historical texts is presented and discussed. The goal of the semantic digitization is to contribute to the study of historical texts, in particular of historical reference works, by making a distinction between the planning of the 'first' and the planning of the 'second', or semantic, retrodigitization. We aim to formulate suggestions for the systematic unfolding (*Ger.* Erschließung) of the semantic level of digitized historical texts.

Literatur

Carriero, Valentina Anita / Gangemi, Aldo / Mancinelli, Maria Letizia / Nuzzolese, Andrea Giovanni / Presutti, Valentina / Veninata, Chiara (2021): "Pattern-based Design Applied to Cultural Heritage Knowledge Graphs", in: *Semantic Web* 12: 313 – 357. 10.3233/SW-200422

Denzel, Markus A. (2002): "Handelspraktiken als wirtschaftshistorische Quellengattung vom Mittelalter bis in das frühe 20. Jahrhundert. Eine Einführung" in: Denzel, Markus A. / Hocquet, Jean-Claude / Witthöft, Harald (eds.): *Kaufmannsbücher und Handelspraktiken vom Spätmittelalter bis zum beginnenden 20. Jahrhundert — Merchant's Books and Mercantile Pratiche from the Late Middle Ages to the Beginning of the 20th Century*. Stuttgart: Steiner Verlag 11-45.

Kendall, Elisa F. / McGuinness, Deborah L. (2019): *Ontology engineering.* (= Synthesis Lectures on The Semantic Web: Theory and Technology, Lecture 18). [California]: Morgan and Claypool. 10.2200/S00834ED1V01Y201802WBE018

Kramper, Peter (2019). *The Battle of the Standards. Messen, Zählen und Wiegen in Westeuropa, 1660-1914*. Berlin / Boston: De Gruyter.

Lodi, Giorgia / Asprino, Luigi / Nuzzolese, Andrea Giovanni / Presutti, Valentina / Gangemi, Aldo / Recupero, Diego Reforgiato / Veninata, Chiara / Orsini, Annarita (2017): "Semantic Web for Cultural Heritage Valorisation", in: Hai-Jew, Shalin (ed.): *Data Analytics in Digital Humanities*. Multimedia Systems and Applications. Springer: Cham 3-37. https://doi.org/10.1007/978-3-319-54499-1_1

Martín-Recuerda, Francisco / Walther, Dirk / Eisinger, Siegfried / Moore, Graham / Andersen, Petter / Opdahl, Per-Olav / Hella, Lillian (2020): "Revisiting Ontologies of Units of Measure for Harmonising Quantity Values – A Use Case", in: Pan, Jeff Z. / Tamma, Valentina / d'Amato, Claudia / Janowicz, Krzysztof / Fu, Bo / Polleres, Axel / Seneviratne, Oshani / Kagal, Lalana (eds.): *The Semantic Web – ISWC 2020*. (= Lecture

Notes in Computer Science, vol. 12507). Springer: Cham 551-567. https://doi.org/10.1007/978-3-030-62466-8_34

Noback, Christian / Noback, Friedrich (1850): *Vollständiges Taschenbuch der Münz-, Maass-, und Gewichtsverhältnisse, der Staatspapiere, des Wechsels- und Bankwesens, und der Usanzen aller Länder und Handelsplätze*. Leipzig: F.A. Brockhaus.

Paju, Petri / Oiva, Mila / Fridlund, Mats (2020): "Digital and distant histories: Emergent approaches within the new digital history", Fridlund, Mats / Oiva, Mila / Paju, Petri (eds.): *Digital histories: Emergent approaches within the new digital history*. Helsinki: Helsinki University Press 3-18. https://doi.org/10.33134/HUP-5-1

Witthöft, Harald (2018): "Numerical Communication in Intercontinental Trade and Monetary Matters: Coins and Weights in China and East Asia in Merchants' Pocketbooks and Commercial Guides (16th–19th Centuries)", in: Theobald, Ulrich / Cao, Jin (eds.): *Southwest China in a Regional and Global Perspective (c. 1600-1911)*. Leiden / Boston 225-290. https://doi.org/10.1163/9789004353718_009