Intra-linking the Research Corpus: Using Semantic MediaWiki as a lightweight Virtual Research Environment

Schindler, Christoph
schindler@dipf.de
German Institute for International Educational Research, Germany

Ell, Basil
basil.ell@kit.edu
Karlsruhe Institute of Technology, Germany

Rittberger, Marc
rittberger@dipf.de
German Institute for International Educational Research, Germany

In recent years, virtual research environments (VREs) emerged as a topic referring to the established research field in the digital humanities: enabling research practices with digital tools. First projects in this area are realized and discussed by the community (Carusi 2010; Dunn 2009; Neuroth et al. 2009). In the humanities, researchers point out that the so-called ‘data deluge’ (Hey et al. 2003), which has influenced several national and supranational information policy agendas in the sciences, does not cover the full range of aspects of research practices in the humanities. While digital libraries and archives offer a new plurality of research resources in the humanities, the ‘complexity deluge’ (Dunn 2009) formulates an opposite agenda addressing the sometimes fuzzy, interfering and dispersed practices of humanities research. In this paper we want to address this tension between research data, metadata and collaborative action in the design of research corpora carried out within a VRE. Therefore we will focus the ongoing corpus re-arrangement and the potentials of Social Semantic Media technologies to expand metadata creation and use in qualitative research. Tools for qualitative research target the flexible coding system, i.e. allowing researchers to annotate research resources according to a classification system that may evolve over time. However, the tools have been criticized for the limited metadata interoperability of resources and research findings (Corti et al. 2011). In our paper, we outline the aspects of interoperability in qualitative corpora research and focus on researchers’ capabilities to intra-link the corpus. We use the term intra-linking to address
a main aspect in qualitative research: to create and
to describe entities while allowing for the ongoing re-arrangement of entities and their properties in the
research process. These capabilities will be
discussed and exemplified within the scope of
the project Semantic MediaWiki for Collaborative
Corpora Analysis (SMW-CorA) which aims to
reconfigure Semantic MediaWiki as a lightweight
virtual research environment.3

The field of corpora centered research in the digital
humanities offers interesting insights into the design
of VREs. In the early 1990s, Biber pointed out
main aspects of corpus design by problematizing a
priori determinations of its boundaries and formal
specifications. He recommends the selection of
relevant objects and the formal description to be
realized as a cyclic or iterative process of corpus
work (Biber 1993: 256). While a linguistic approach
mainly aims at a statistical ‘representation’ in
relation to a target population, qualitative corpus
research, which is focused here, pursues a so
called qualitative selection, i.e. a typification of yet
unknown properties in research (Bauer 2000: 20).
We argue that this indeterminacy of entities and
properties in qualitative research emphasizes the
affordance of a VRE enabling researchers to intra-
link the corpus – it means to give them the ongoing
capabilities to create, modify and re-arrange entities
and properties while doing research. This topic of
qualitative corpus research addresses the research
and design desideratum of qualitative annotations
(Juola 2008) and a demanded shift to further
capabilities for the researcher to control the data
(Smith 2008: 178).

While the SMW-CorA project targets the re-use of
its VRE infrastructure in different research contexts
in the mid-term, its initial design is subject to a
cooperation with a research project in the history
of education, involving a major library in this field.
The educational research project encompasses the
analysis of a corpus of 25 educational lexica dating
from 1774 to 1942, reconstructing the development of
educational science. Discourse, field and content
analysis are supported and applied to grasp the
networked relationships in this scholarly field.
The Library for the History of Education (BBF)
hosts a large part of the lexica at the digital
library Scripta Paedagogica Online (SPO).4 The
collection amounts to nearly 22,000 articles. Each
lexicon is bibliographically described and accessible
online as image files. Participatory and agile design
approaches are used to offer an adequate shared
space for the different stakeholders to articulate
possible potentials and boundaries of the VRE.

The SMW-CorA project builds on a Social Semantic
Media technology which in turn is based on the Web
2.0 software MediaWiki (MW),5 which is used at
the well-known online encyclopedia Wikipedia, and
the extension Semantic Media Wiki (SMW).6 The
latter enables the use of semantic annotations and
integration within the Semantic Web through import
and export of semantic data and linking to external
entities. While other VREs have been realized
using this wiki technology, the SMW-CorA project
promotes research on a corpus by configuring the
facilities of MW and SMW and by developing further
extensions to offer a configurable and lightweight
VRE.

As such, the basic framework MediaWiki offers
some facilities for creating and typifying entities
in a corpus for research (in our case: lemmas,
authors, institutions etc.). The online encyclopedia
Wikipedia demonstrates, besides the collaborative
creation of texts, the capability to store documents
and digital objects. Therein a wiki page can be related
to other entities by hyperlinks and arranged within
a hierarchical category system. SMW extends this by
offering an increased granularity for describing and
linking the corpus by adding metadata descriptions
used in the Semantic Web. The unspcific hyperlinks
between entities can be typified by metadata and
thereby entities can be enriched with attributes or
semantic relations to other entities. Furthermore, it
is possible to import and export metadata in the
Semantic Web standard RDF.

Within the scope of the project a set of use-
cases is explored comprising the scholarly work
in the research life cycle from importing research
resources, coding, classifying and analyzing these to
the export for re-use. The supported and envisaged
use cases have in common that entities can be
integrated, created, modified and interlinked (i.e.
intra-linking). This functionality is focused within
the project and supported by tools to enable
researchers to carry out an ongoing re-arrangement
of their corpora. While the aspects of importing
lexica from a digital library and of exporting the
bibliographic data in RDF are discussed in (Schindler
et al. 2011), further use cases exemplify these
capabilities. Besides this import functionality, the
enrichment of entities such as editors, authors, and
related affiliations with properties by using Semantic
Web Browsing technologies is a further example.8
Thereby, semantic properties from digital archives,
libraries or further collections can be integrated
and adapted to the locally used metadata schema
in the VRE. This enables researchers for example
to add biographical data of authors or editors
from authority files (e.g. German National Library
GND). Furthermore the import and collaborative
development of taxonomies (e.g. classification or
coding schema) are supported by interlinking entities
within the VRE and linking entities into the Web of
Data. These links to the world external to the corpus, together with the data export facilities of the VRE, enable the reuse of the content created within the VRE. Additionally, the VRE provides functionalities for creating and re-arranging metadata schema as well as a bottom-up task management to allow supervision of the research process.

To summarize, we identified the need of researchers to create, manage and intra-link entities and metadata objects in a research corpus. This functionality is relevant for multiple use cases where researchers perform a qualitative analysis on a corpus of resources such as digital/digitized documents and images. Our main contribution is the development of a lightweight collaborative and adaptive VRE that enables researchers to perform these tasks as well as to enable export of the created data and the content’s sharing and reuse. Since the VRE is based on a flexible Open Source platform it can be tailored by the researchers towards their specific needs. Therefore this lightweight environment may serve as a starting point for further re-uses and re-configurations in unforeseen research settings and functionalities required in the future.

Funding
This work was supported by the German Research Foundation (DFG) in the domain of ‘Scientific Library Services and Information Systems’ (LIS). The funded project is entitled: ‘Entwicklung einer Virtuellen Forschungsübung für die Historische Bildungsforschung mit Semantischer Wiki-Technologie – Semantic MediaWiki for Collaborative Corpora Analysis’ [INST 367/5-1, INST 5580/1-1]

References


Notes
1. It should be mentioned that a similar field of tension is articulated for the sciences as ‘science friction’ (Edwards et al. 2011) by addressing the problems of different disciplines working on the same phenomena and trying to interoperate.

2. This term refers to the concept of ‘intra-action’, which describes the interrelations and re-configurations of research apparatuses in respect of ‘locally stabilized phenomena’ (Barad 2003: 817).
Corpus Coranicum: A digital landscape for the study of the Qu’ran

Schnöpf, Markus
schnoepf@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

1. Introduction

In 2007 begun the project Corpus Coranicum located at the Berlin-Brandenburg Academy of Science and Humanities with an estimated duration of twelve years. A goal of the project is a holistic documentation of the holy text. The project consists of different modules: Collection of early manuscripts, documentation of environmental texts to the Quran, documentation of alternate writings and finally a commentary on each sura of the Quran. In the last years the technological infrastructure has been set up and data was collected in a SQL database. The commentary of the was realised in XML and is stored in a XML-database. The website of the project thus combines SQL and XML in an integrated information system. Lately, a bibliography consisting of 8000 references was added to the system. Further investigations are directed more on a scientific dating, analysing the materiality of early written documents. They are scheduled for 2012/13 within a French-German joint research project. Another new module is a glossary of the and early Arabic literature. For overcoming troubles in the presentation of early Arabic script a special font has been developed: The Coranica.

The project Corpus Coranicum began at the Berlin-Brandenburg Academy of Science and Humanities in 2007 with a planned duration of 12 years. The project aims at both a holistic edition of the Quran and also an extensive commentary. In addition to the Al-Azhar Quran edition from 1923/1924, this project will provide the reader with early written testimonies as well as oral reading variants that are manifested in early Islamic literature.

2. Manuscripta

The project sees the module Manuscripta Coranica as following the tradition of G. Bergsträbers planned Apparatus Criticus, which due to Bergsträbers early death could never be realised. Thus the project aims at a new approach to the text of the Quran. Bergsträber collected the oldest Quran manuscripts