Linguistics and Literature Studies 5(2): 99-104, 2017
DOI: 10.13189/lls.2017.050205 http://www.hrpub.org

Integrating Canonical Text Services into CLARIN's Search Infrastructure

Jochen Tiepmar*, **Thomas Eckart**, **Dirk Goldhahn**, **Christoph Kuras**

Department of Natural Language Processing, Faculty of Mathematics and Computer Science, University of Leipzig, Saxony, Germany

Copyright ©2017 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Today's digital research infrastructures target a variety of user groups. A key task to achieve acceptance and active participation among them are both user-friendly and machine-readable interfaces to digital resources. This is especially the case for highly integrated infrastructures like the CLARIN project. The Canonical Text Service Protocol CTS is an established system in document based Digital Humanities that covers many of associated problems, like dealing with varying levels of text granularity, persistent identification, address resolution and simple interfaces for an integration in various automatic work flows. The paper shows the advantages of integrating a CTS instance into CLARIN and also demonstrates additional benefits of this CTS implementation in form of built-in text mining techniques.

Keywords Canonical Text Service, CLARIN, Linguistic Infrastructures, Webservice, Text Data, Text Mining

1 Introduction

The current landscape of digital resources in the field of humanities can be characterized as rather scattered and oriented on highly specific research interests. Despite strong efforts in building digital research infrastructures (like CLARIN, DARIAH etc.) to overcome the current heterogeneity and to build integrated research environments, it can be assumed that the majority of textual resources in this field (although often being encoded based on standard formats like the TEI guidelines [13]) are still not available via standardized interfaces and can not be found by means of existing search functionality. Naturally, it is a key task to convince and motivate researchers from a wide variety of subfields of the humanities to provide their valuable data to the wider community. As a consequence several attempts have been made and are actively used to minimize the effort needed for a thorough integration in existing environments and work flows, and to provide obvious benefits as motivation for the interested resource provider. The following issues are considered as especially problematic and relevant to the authors and are addressed in this paper:

- Many of the current solutions treat textual resources as atomic, i.e. all provided interfaces[1] are focused on the complete resource. The inherent structure of textual data is left to be processed by external tools or manually extracted by the user. Although this being acceptable for some use cases, a highly integrated research environment loses much of its power and applicability for research questions if ignoring this obvious fact.
- Textual resources do not have a typical granularity. Even for rather similar textual resources (like Web-based corpora or document-centric collections) it can not be assumed to have a "default structure" on which analysis or resource aggregation can take place. As a consequence many approaches require and assume a standard format that is foundation for all provided applications and interfaces.
- Granularity has to be addressed as a basic feature of (almost) all textual resources. Current infrastructures make use of several identification and resolving systems (like Handle, DOI, URNs etc.) but a fine-grained identification and retrieval of (almost) arbitrary parts are hardly supported or have to be modeled artificially using features that these systems provide[2]. As a consequence even textual resources already provided in CLARIN are often not directly accessible or combinable because of the heterogeneity of used reference solutions or the level of supported granularity.

2 Canonical Text Services

The Canonical Text Service protocol [11] describes a framework for web based identification and retrieval of passages of text cited by canonical reference as typical in

[1] Regardless of being focused on direct human interaction or automatic analysis procedures.

[2] For a concrete example using "part identifiers" of the Handle System see [2].

classical studies and other literary disciplines. To achieve this it uses persistent CTS URNs to reference specific text passages such as a complete document or a more specific text part. The following example illustrates this reference system. The CTS URN *urn:cts:pbc:bible.parallel.eng.kingjames:* specifies the complete edition/document "King James Version of the Christian Bible". This URN corresponds to the reference that is provided by alternative systems that work with complete documents. In contrast to those, CTS allows to reference any static text part with URNs like *urn:cts:pbc:bible.parallel.eng.kingjames:1*[3], which references the first book in this document, or *urn:cts:pbc:bible.parallel.eng.kingjames:1.4.2*[4], which references the second sentence of the fourth chapter of the first book of this Bible translation. Static URNs generally point to structural elements of the texts, like chapter, paragraph, sentence or stanza but are not restricted to a specific schema. Additionally CTS URNs allow to reference spans between two static URNs like *urn:cts:pbc:bible.parallel.eng.kingjames:1-1.4.2*[5]. Using sub passage notation, any possible text passage can be requested, like *urn:cts:pbc:bible.parallel.eng.kingjames:1@the[2]-2.4.2@a[3]*[6] - which translates to the text passage from the second "the" in book one to the third character "a" in the second sentence in the fourth chapter of the second book. Sub passages are resolved using the exact string, which normally correlates to a word but may also be a string of words or a letter. For the remainder of this paper only the static CTS URNs are of relevance, URNs for spans of text passages and sub passage notation are not included. The relationship between CTS URNs and different levels of text granularity is also depicted in figure 1 for an excerpt of the English King James translation of the Bible.

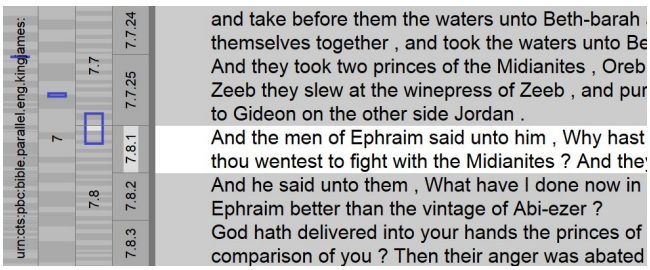

Figure 1. Hierarchical mapping between CTS URN references and text elements

One central aspect of CTS is that it was developed by humanists and reflects their perspective on text references, as for instance described by [3]. This perspective might differ from the one used in other communities. Specifically much semantics is implicitly or explicitly encoded in CTS URNs, which is different from the understanding that IDs should be opaque as it is often assumed in CLARIN. Combining CTS with CLARIN has the potential to create an important connection between the two philosophies. It opens up tools that are provided in CLARIN for the digital classicist community and also provides access to their valuable data. The used implementation of CTS (described in [12]) proved to be efficient and scalable even for large text collections. Additionally it became possible to implement several features that are not part of the CTS protocol but useful additions that rely on certain properties of CTS, like text structure based text alignment (see section 7). The data sets that are already available as instances of CTS - and therefore can be imported into CLARIN with references to individual text parts - include Perseus[7], Parallel Bible Corpus[8], the Deutsche Textarchiv[9] and many others. Even if the data is already included in CLARIN - like the documents of the Deutsches Textarchiv - importing them via CTS provides resources with a smaller granularity and therefore additional research value. New data sets can be imported from TEI/XML in a configurable workflow or created with project specific import scripts from any format. For this work a subset of the Parallel Bible Corpus is used as an example. This subset consists of 20 Bible translations that contain more than 30,000 text parts and are either older than 70 years or published as public domain.

3 CTS as a Uniform Text Communication Interface

The following example illustrates the usefulness of integrated support for CTS using Monica Berti's Digital Athenaeus Index Digger[10]([1]) as an example. The columns

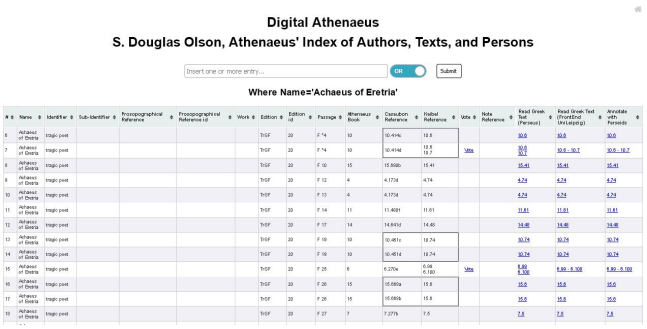

Figure 2. Digital Athenaeus Index Digger

Read Greek Text (Perseus), *Read Greek Text (Frontend UniLeipzig)* and *Annotate with Perseids* provide hyperlinks to different tools that work with different data sets and are developed by different research groups. Yet, because CTS is specified as a technology independent protocol, it is possible to link the same text passages together by providing the

[3] http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1

[4] GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.4.2

[5] GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1-1.4.2

[6] GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1@the[2]-2.4.2@a[3]

[7] http://www.perseus.tufts.edu/hopper/

[8] http://paralleltext.info/data/

[9] http://www.deutschestextarchiv.de/

[10] http://www.digitalathenaeus.org/tools/OlsonIndex/digger.php?what[]=Name—Achaeus+of+Eretria&onoffswitch=on

same CTS URN as a parameter in the hyperlink. This makes it possible to connect and combine results of different tools and significantly increase interoperability between research groups. This also reduces the amount of effort that has to be put into retrieval of text resources and format conversions for new research projects.

4 The CLARIN Infrastructure

CLARIN[11] is the short name for the "Common Language Resources and Technology Infrastructure", a research infrastructure for scholars in the humanities and social sciences ([8]). It offers easy and sustainable access to digital language data (e.g. in written, spoken or multimodal form) and also advanced tools to discover, explore, exploit, annotate or analyse data sets. CLARIN follows the approach that all integrated digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the respective fields. Therefore, a networked federation of language data repositories, service and knowledge centers, with access for all members of the academic communities in all participating countries was created. In CLARIN tools and data from the different centres are interoperable, so that they can be combined to perform complex operations in order to support researchers in their work.

CLARIN is one of the research infrastructures of the European Research Infrastructures Roadmap by ESFRI, the European Strategy Forum on Research Infrastructures. This roadmap contains five research infrastructures in the area of social sciences (CESSDA, European Social Survey, and SHARE) and humanities (CLARIN and DARIAH). On the European level CLARIN's governance and coordination body is CLARIN ERIC. An ERIC is an international legal entity, established by the European Commission in 2009. In 2012 CLARIN ERIC was established and took up the vital mission to develop and maintain this international infrastructure.

By now, the CLARIN infrastructure is fully operational in many countries, among them the German consortium CLARIN-D[12]. Therefore, a large number of participating centers are offering access to data, tools and expertise. At the same time, CLARIN continues being established in countries that joined just recently, and CLARIN's datasets and services are constantly updated, extended and improved.

5 CLARIN Integration

For a thorough integration in CLARIN the granularity of text resources contained in CTS instances has to be exposed mainly via metadata, as the comparable interfaces for retrieval of the actual text material is already provided by every CTS instance.

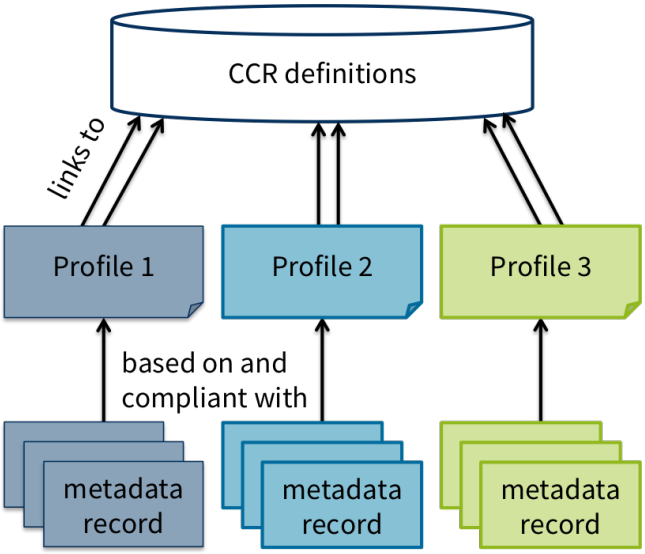

Figure 3. General scheme of component-based metadata according to CMDI [5]

For describing all kinds of resources CLARIN makes use of the Component Metadata Infrastructure (CMDI) framework [5]. It allows to build and use component-based metadata schemata for all kinds of resources including descriptions of complete corpora, tools or services. For the concrete realisation the popular CMD profile "OLAC-DcmiTerms" (clarin.eu:cr1:p_1288172614026) was used. The REST-based design of the CTS protocol and its implementations reduce the effort that is necessary to include CTS instances in the center-based CLARIN infrastructure. The CLARIN center Leipzig makes strong use of webservices for both its internal structure and the external interfaces it provides[13]. For the incorporation of potentially unlimited numbers of CTS instances this approach was extended by creating a wrapper webservice as the main interface for the internal center infrastructure. Regarding all metadata-centric external views the default repository system is still used and provides a transparent interface to the CTS resources by standard interfaces like OAI-PMH.

The implemented solution allows the creation of CMD-compliant metadata on every potential level of granularity that is provided by the CTS instance. For the time being it was decided to only expose the top two levels via metadata. For the example depicted in section 2 it means that every specific edition of a Bible and all of its books are described by their own metadata files and can be accessed and searched for in search engines. This includes the typical descriptive metadata, all relevant references to resource-specific CTS services and the hierarchical interlinkage of metadata files as it is supported by the Virtual Language Observatory [4].

[11] https://www.clarin.eu
[12] https://www.clarin-d.de
[13] For details see [2].

6 Results

The result of the process is a collection of metadata files that are created using the meta information that is part of the text inventory of the CTS instance. These files can be accessed using the Virtual Language Observatory[14]. As an example, the textgroup *urn:cts:bible* is translated into a single CMD record and the same holds for each document level CTS URN in this textgroup, like *urn:cts:pbc:bible.parallel.arb.norm:*, *urn:cts:pbc:bible.parallel.ceb.norm:* and *urn:cts:pbc:bible.parallel.ces.norm:*.

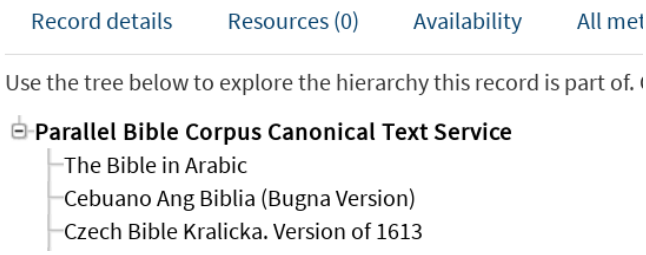

Figure 4. Visualisation of a CTS resource in the VLO

For each document the CTS request for the CTS URNs on citation level 1 - which corresponds to one book of the Bible like Genesis - is added as a resource along with the CTS request for the complete document and the TEI/XML source file that was imported in CTS. The citation level can be chosen arbitrarily. In this work it seemed to be reasonable to work with 60 to 70 references per document to the books of the Bible instead of more than 1000 references per document on citation level 2 or even more than 30 000 references per document on citation level 3.

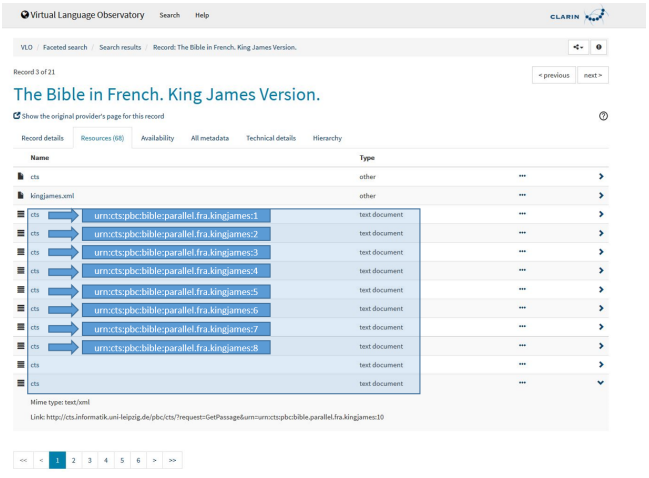

Figure 5. CTS requests as external resources of a CMD record

7 CTS Applications

Several applications were developed that rely on properties of the described implementation. Using CTS as a standardized input they can be used with any data that is stored in the system. Each of these applications is provided along with the CTS implementation and can be used as soon as a new server instance is created. It is especially not required to pre-calculate any additional data.

One of these applications is the Candidate Text Alignment Browser[15] that visualises textual variants in one language using the TRAViz library [9].

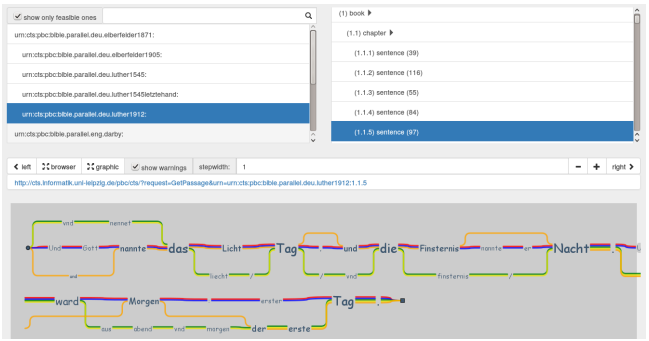

Figure 6. Candidate Text Alignment: align text variants in one language

The tool allows users to align parts of a document with their counterparts in other editions. A document is a candidate for such an alignment if only the last part of its URN differs from the input URN. As the output all variations of the selected text in the candidate set are visualized and differences between all selected editions are illustrated. Figure 7 contains a specific example for variations in Genesis 1.5 ("God called the light Day, and the darkness he called Night.") in several editions of the Bible in German language. As one major benefit for users of the Digital Classics this visualization intuitively reflects diachronic changes in biblical texts over almost 400 years. Further parameterization allows to change the size of the text passage, although, since TRAViz is relatively memory intensive, it is not recommended to use elongated passages as input.

The CTS server also integrates the Parallel Alignment Browser[16] that can be used to align text passages from selected documents independent of their language.

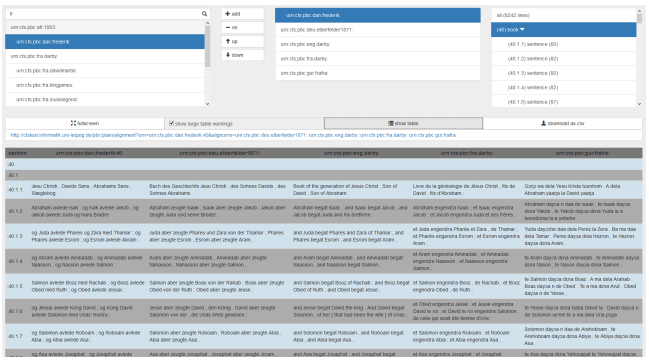

Figure 7. Parallel Text Alignment: align text variants in several languages

[14] https://vlo.clarin.eu

[15] http://cts.informatik.uni-leipzig.de/cts_admin_tools/alignbrowser/?ctsURL=../../pbc

[16] /parallelbrowser/?ctsURL=../../pbc&sep=:

This can, for example, be used to spot structural variations in different translations. The results are visualized as a table and can be exported in standard file formats. The number of documents the analysis can be applied to is only limited by the number of documents in the CTS server.

For better readability or easier creation of CTS URNs the Canonical Text Reader and Citation Exporter CTRaCE[17] was developed. This tool renders the output of a CTS instance in a more appealing way, lets users traverse the documents and easily create CTS URNs for a selected text passage and generally provides "an interface to intuitively make this [meaning:CTS] capability accessible to humanities scholars."([10]).

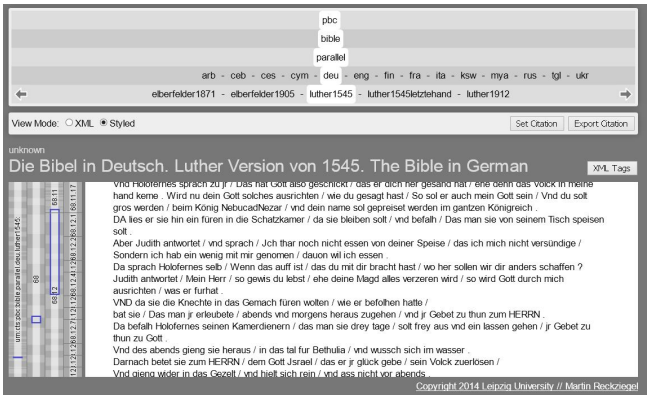

Figure 8. CTRaCE: Canonical Text Reader and Citation Exporter

Integrating support for Canonical Text Services in CLARIN makes it possible to connect these and future applications with the numerous tools and methods already provided in the CLARIN infrastructure.

8 Future Work

The current setup of Leipzig's CTS server is fully functional and will be a valuable part of CLARIN's constantly growing resource landscape. The integration of more resources in the CTS instance is already in progress and the described work also significantly simplifies the inclusion of CTS instances hosted by other data providers in CLARIN.

For an even tighter integration the current focus of development lies on providing interfaces to more relevant CLARIN components. It is expected that especially the preparation of an endpoint compliant to the CLARIN Federated Content Search FCS[18] or a wrapper for the execution environment WebLicht [7] will boost the usefulness of the system. Furthermore, continuous efforts are being made to provide more integrated analysis and visualization components in the CTS server.

[17] /browser /?ctsURL=../../pbc/cts/

[18] http://weblicht.sfs.uni-tuebingen.de/Aggregator/

Acknowledgements

Part of this work was funded by the German Federal Ministry of Education and Research within the project Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig (BMBF 01IS14014B) and CLARIN-D (BMBF 01UG1120C).

REFERENCES

[1] M. Berti, C.W. Blackwell, M. Daniels, S. Strickland, and K. Vincent-Dobbins. Documenting Homeric Text-Reuse in the Deipnosophistae of Athenaeus of Naucratis, Digital Approaches and the Ancient World. Ed. by G. Bodard, Y. Broux, and S. Tarte. BICS Themed Issue 59(2), 2016, 121-139

[2] Volker Boehlke, Torsten Compart, and Thomas Eckart. Building up a CLARIN resource center – Step 1: Providing metadata, Workshop on Describing Language Resources with Metadata, LREC, Istanbul, 2012.

[3] Gregory Crane, Bridget Almas, Alison Babeu, Lisa Cerrato, Matthew Harrington, David Bamman, and Harry Diakoff. Student researchers, citizen scholars and the trillion word library, Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 213–222, 2012.

[4] Twan Goosen and Thomas Eckart. Virtual Language Observatory 3.0: What's New?, CLARIN annual conference 2014 in Soesterberg, The Netherlands, 2014.

[5] Twan Goosen, Menzo Windhouwer, Oddrun Ohren, Axel Herold, Thomas Eckart, Matej Durco and Oliver Schonefeld. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure, Selected Papers from the CLARIN 2014 Conference. 2015.

[6] Gerhard Heyer, Thomas Eckart, and Dirk Goldhahn. Was sind IT-basierte Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften und wie können sie genutzt werden?, Information - Wissenschaft und Praxis, Volume 66, S. 295-303, De Gruyter, 2015.

[7] Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. WebLicht: web-based LRT services for German, Proceedings of the ACL 2010 System Demonstrations, S. 25–29. Association for Computational Linguistics, 2010.

[8] Erhard Hinrichs and Steven Krauwer. The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), May 2014, 1525–31.

[9] Stefan Jänicke, Anette Geßner, Marco Büchler, and Gerik Scheuermann. Visualizations for Text Re-use, Proceedings of the 5th International Conference on Information Visualization Theory and Applications, IVAPP 2014, pages 59–70, 2014.

[10] Martin Reckziegel, Stefan Jänicke, and Gerik Scheuermann. CTRaCE: Canonical Text Reader and Citation Exporter, Proceedings of the Digital Humanities 2016, Krakow, 2016.

[11] David Neel Smith. Citation in classical studies, Digital Humanities Quarterly, 3. 2009.

[12] Jochen Tiepmar. Release of the MySQL-based implementation of the CTS protocol, 3rd Workshop on the Challenges in the Management of Large Corpora, 2015.

[13] TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium., http://www.tei-c.org/Guidelines/P5/, accessed: 20161128.