Canonical Text Services in CLARIN - Reaching out to the Digital Classics and beyond

Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn, Christoph Kuras
Natural Language Processing Group
University of Leipzig, Germany
{jtiepmar,teckart,dgoldhahn,ckuras}@informatik.uni-leipzig.de

Abstract

Providing both user-friendly and machine-readable interfaces to digital resources is one of the key tasks of highly integrated research infrastructures like CLARIN. The presented implementation of the Canonical Text Service Protocol CTS covers many of the associated problems, like dealing with varying levels of text granularity, persistent identification and address resolution, and simple interfaces for an integration in various automatic workflows. The paper also demonstrates additional benefits of our CTS implementation in form of built-in text mining techniques.

1 Introduction

The current landscape of digital resources in the field of humanities can be characterised 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 being often encoded based on standard formats like the TEI guidelines) are still not available via standardised 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 workflows, 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 usage of several identification and resolving systems (like Handle,

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

DOI, URNs etc.) but a fine-grained identification and retrieval of (almost) arbitrary parts are hardly supported or have to be artificially modelled 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.

The authors assume that these problems (among others) have to be solved to provide more useful applications, services and workflows to the interested users, and to have a solid foundation for a highly integrated and appealing research infrastructure (Heyer et al., 2015). In the following, an implementation of the Canonical Text Services (CTS) will be introduced. Its integration into the CLARIN infrastructure will be discussed and its contributions to overcoming the problems described will be clarified.

2 CTS

Canonical Text Service by Smith (2009) is a framework for web based identification and retrieval of passages of text cited by canonical reference as typical in 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 the benefits of 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 level of text granularity is also depicted in figure 1 for an excerpt of the book of Genesis ("and the rib, which the Lord God had taken from man, made he a woman, and brought her unto the man.").

One central aspect of CTS is that it was developed by humanists and reflects their perspective on text references, as for instance described by Crane et al. (2012). 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.

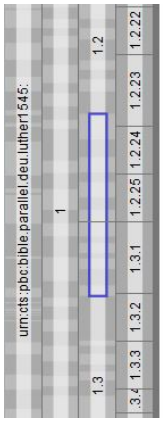
[2] For a concrete example using "part identifiers" of the Handle System see (Boehlke et al., 2012).

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

[4] `http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.4.2`

[5] `http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1-1.4.2`

[6] `http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1@the[2]-2.4.2@a[3]`



Vnd Gott der HERR bawet ein Weib aus der Riebe / die er von dem Menschen nam / vnd bracht sie zu jm .
Da sprach der Mensch / Das ist doch Bein von meinen Beinen / vnd Fleisch von meinem fleisch / Man wird sie Mennin heissen / darumb / das sie vom Manne genomen ist .
Darumb / wird ein Man seinen Vater vnd seine Mutter verlassen / vnd an seinem Weibe hangen vnd sie werden sein ein Fleisch .
Vnd sie waren beide nacket / der Mensch vnd sein Weib / vnd schemeten sich nicht (Jd est) / Dürfften sich nicht schemen .
VND die Schlange war listiger denn alle Thier auff dem felde / die Gott der HERR gemacht hatte / vnd sprach zu dem Weibe / Ja / solt Gott gesagt haben / Jr solt nicht essen von allerley Bewme im Garten ?
DA sprach das Weib zu der Schlangen / Wir essen von den früchten der bewme im Garten .
Aber von den früchten des Bawms mitten im Garten hat Gott gesagt / Esset nicht da von / rürets auch nicht an / Das jr nicht sterbet .
Da sprach die Schlang zum Weibe / Jr werdet mit nicht des tods sterben /

Figure 1: Relationship between CTS URNs and text passages in a bible of 1545

The used implementation of CTS (described in (Tiepmar, 2015)) 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 4). 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], Deutsches Textarchiv[9] (DTA) and many others. Even if the data is already included in CLARIN - like the documents of the DTA - 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 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. 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[10]. For the incorporation of potentially unlimited numbers of CTS instances this approach was extended by creating a wrapper webservice as 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 that means that every specific edition of a bible and all of its books are described by their own

[7] `http://www.perseus.tufts.edu/hopper/`
[8] `http://paralleltext.info/data/`
[9] `http://www.deutschestextarchiv.de/`
[10] For details see (Boehlke et al., 2012).

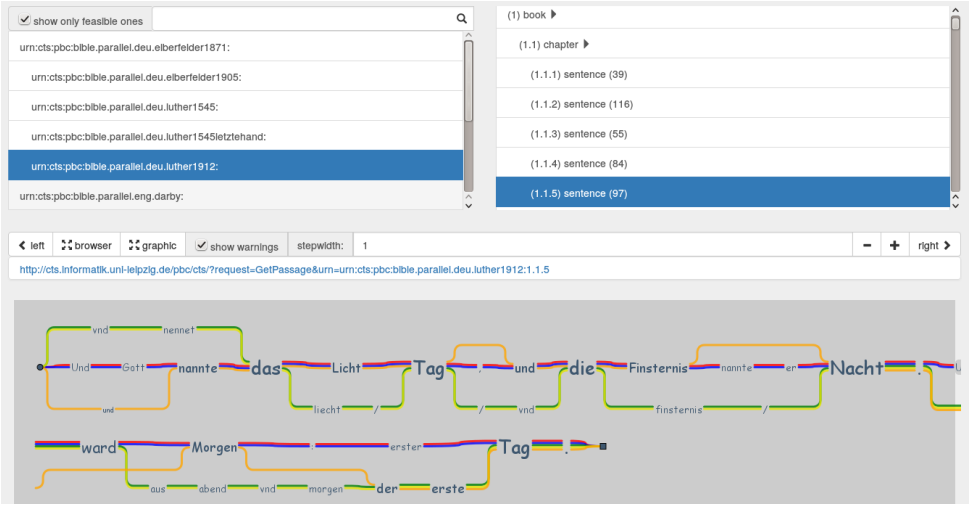


Figure 2: Candidate Text Alignment: align text variants in one language

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 (Goosen and Eckart, 2014).

4 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[11] that visualizes textual variants in one language using the TRAViz library (Jänicke et al., 2014).

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 output all variations of the selected text in the candidate set are visualised and differences between all selected editions are illustrated. Figure 2[12] 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 visualisation intuitively reflects diachronic changes in biblical texts over almost 400 years. Further parameterisation 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[13] that can be used to align text passages from selected documents independent of their language. 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

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

[12] http://cts.informatik.uni-leipzig.de/cts_admin_tools/alignbrowser/?ctsURL=../../pbc&urn=urn:cts:pbc:bible.parallel.deu.luther1912:1.1.5&stepwidth=1

[13] http://cts.informatik.uni-leipzig.de/cts_admin_tools/parallelbrowser/?ctsURL=../../pbc&sep=:

Exporter CTRaCE[14] was developed by Reckziegel et al. (2016). These tools render the output of a CTS instance in a more appealing way, let users traverse through the documents and easily create CTS URNs for a selected text passage.

5 Further work

The current setup of Leipzig's CTS server is already fully functional and will be a valuable part of CLARIN's constantly growing resource landscape. 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[15] or a wrapper for the execution environment WebLicht (Hinrichs et al., 2010) will boost the usefulness of the system. Furthermore, continuous efforts are being made to provide more integrated analysis and visualisation components in the CTS server.

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).

References

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

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

[Goosen and Eckart2014] Twan Goosen and Thomas Eckart. *Virtual Language Observatory 3.0: What's New?* CLARIN annual conference 2014 in Soesterberg, The Netherlands.

[Heyer et al.2015] Gerhard Heyer, Thomas Eckart, and Dirk Goldhahn. 2015. *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.

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

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

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

[Smith2009] David Neel Smith. 2009. *Citation in classical studies.* Digital Humanities Quarterly, 3.

[Tiepmar2015] Jochen Tiepmar. 2015. *Release of the MySQL-based implementation of the CTS protocol.* 3rd Workshop on the Challenges in the Management of Large Corpora.

[14] `http://cts.informatik.uni-leipzig.de/cts_admin_tools/browser/?ctsURL=../../pbc/cts/`
[15] `http://weblicht.sfs.uni-tuebingen.de/Aggregator/`