Data repositories in the Humanities and the Semantic Web: modelling, linking, visualising

Max Grüntgens and Torsten Schrade

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2, 55131 Mainz, Germany
{max.gruentgens,torsten.schrade}@adwmainz.de
http://www.adwmainz.de

Abstract. The paper discusses the inherent potential of the Semantic Web and its related technologies for humanities research. The focal point lies on the extraction of semantic relations from heterogeneous XML based scholarly corpora using a webservice based infrastructure (XTriples). Especially the creation of methodologically distinct semantic corpora stemming from data sets originating in the humanities will be discussed. During this discussion, questions of modelling, linking, and visualising data from the humanities will be tackled as well.
Finally, opportunities for further analysis and visualisation of semantically modelled data in the humanities are exemplarily presented and discussed.

Keywords: Semantic Web, XML, RDF, scholarly editing, data repositories

1 The Semantics of XML-based Data Repositories in the Humanities

Many data repositories in the humanities base their data modelling on the *Text Encoding Initiative* (TEI) and use XML as their primary data format. XML is well suited for digital scholarly editions and the philological and editorial tasks associated with this field of historical research, because it fulfills the important criteria of interoperability, sustainability, and reusability. By applying standards-compliant TEI markup to the objects of research, they acquire a formal as well as internal structure.

Due to the markup structures used, the data already contains a lot of semantic references, e.g. spatial references, relations between individuals, or conceptual references. From the perspective of the Semantic Web these references are merely implicit in the data (cf. Listing 1). Implicit references have to be transformed into explicit semantic annotations (e.g. RDF) to make the data usable for Semantic Web approaches (see also [14]).

Listing 1: Implicit Semantics (TEI-XML)

```
<correspDesc key="686" cs:source="#SOE20">
    <correspAction type="sent">
        <persName ref="http://d-nb.info/gnd/118540238">
            Johann Wolfgang von Goethe
        </persName>
        <placeName ref="http://www.geonames.org/2812482">;
            Weimar
        </placeName>
        <date when="1793-12-05">
            5.12.1793
        </date>
    </correspAction>
    [...]
</correspDesc>
```

(Inherent semantics in XML: Goethe sends a letter from Weimar 1793. RDF subjects are implicit in the `cs:source` and `ref` attributes. Predicates are implicit in the `date` tag and the `type` attribute. Objects values are implicit in the `correspDesc` tag, the `placeName` and `persName` text nodes and the `when` attribute.)

Most approaches in the humanities base their efforts in this respect on the *Resource Description Framework* (RDF). RDF enables scholars to formulate and annotate semantic statements in a clear and concise way by utilising the triple notation: Subject—Predicate—Object. RDF's particular strong point lies in interlinking, merging and analysing (reasoning) essentially distinct data sets. RDF in general and webservices like *XTriples* in particular can also be used as a means to bridge the gap between so called "altruistic and egoistic modelers" ([2] par. 20) by easiliy providing project specific data in a more abstracted and overarching data format. From a data modelling point of view, RDF has a higher level of abstraction in comparison to data encoded on the basis of TEI-XML (cf. listing 2; see [13]).

Listing 2: Explicit Semantics (RDF)

```
Goethe     is_a              Person  ;
           sends             Letter  .
Letter     dates_to       1793-12-05 ;
           sent_from         Weimar  .
Weimar     is_a              City    ;
           has_longitute     11.32   ;
           has_latitude      50.98   .
```

(The same example as above in simplified `turtle` notation: First column shows subjects, second column shows predicates, third column shows objects)

Although the formal development of structures and metadata within texts and other objects of research in the humanities can be considered rather advanced, the development of semantics and semantic annotations is still lagging

behind. Admittedly, toponyms, personal names or work titles are often annotated nowadays. These annotations, however, confine themselves mostly to pointing out the occurences of a specific entity within a data set.

The annotations used within data repositories in the humanities often fall short of the potential offered by Semantic Web technology in general, and more specifically, by concepts like *linked open data* (LOD). LOD offers the opportunity to join isolated data sets. Thereby, researchers hope to create new, interdisciplinary, and stimulating viewpoints towards established topics.

A key factor in this regard is that LOD and RDF extend established standards from the *Digital Humanities*, e.g. TEI-XML, with additional uses, terminologies, and metadata schemata (see [4]). Due to this extension it is possible to consistently describe and analyse data sets in form and content, regardless if they are of distributed provenience or encompass different internal structures.

Currently there exists a wide gap: On one side of the divide lie the humanities' numerous repositories containing data with great semantic potential, on the other side lie the Semantic Web technologies and data models, which could open up new methods of analysis. Although some concepts, methods, and also tools exist for transforming TEI-XML into RDF and vice versa, these are for the most part overly complex, partially obsolete on a technical level, only implemented as prototypes, or are highly specialised for a specific type of transformation scenario.[1]

Whereas the computer sciences regard the technologies and concepts underlying the Semantic Web as more or less fully developed and applicable (see [10], pp. 11–35), the humanities lag behind in creating representative data sets that demonstrate the usefulness of Semantic Web approaches for their field of academic research.

[1] Projects like *SPQR* (`http://spqr.cerch.kcl.ac.uk/?page_id=3`) or the *Textual Encoding Framework* (`http://rdftef.sourceforge.net/`) are outdated or cannot be generalised. *DERI*'s (`http://www.w3.org/Submission/xsparql-language-specification/`) *XSPARQL Language Specification* can be seen as a highly interesting attempt in this regard. XSPARQL was presented as a W3C Member Submission in 2009. Nevertheless, it still lacks an implementation with a practical orientation. The use of RDFa provides another possibility for semantic markup within an XML file. The data repositories in the humanities which implement RDFa are yet in the minority. The *GRDDL*-Framework (Gleaning Resource Descriptions from Dialects of Languages; `http://www.w3.org/TR/grddl/`), which was established as a W3C Recommendation in 2007, is still a mere theoretical specification. TEI's *OxGarage* webservice (`http://www.tei-c.org/oxgarage/`) provides routines for the conversion from TEI to RDF, but applies only CIDOC-CRM as ontology. Thus *OxGarage* is not capable to utilize and support other ontologies out-of-the-box. In addition the webservice can not include external resources or repositories during processing. Also, it lacks the functionality to return other formats beyond RDF/XML.

2 Creating Semantic Statements from XML Repositories with the XTriples Webservice

The basic principle of creating RDF statements from XML is rather simple. If an XML file's URI or a data unit within this file is regarded as the subject of a triple, then it is possible to assign other data units from the same file or URIs of other resources as objects. Subjects and objects are bound together by predicates from controlled vocabularies. On the whole, the process of translating XML to RDF is therefore mainly focused on the determination of general statement patterns, which can then be applied to and extracted from all resources of the data set in question.

To facilitate this kind of semantic extraction, the *Digital Humanities* departement of the *Academy of Sciences and Literature Mainz* (`www.digitale-akademie.de`) has developed a generic webservice called *XTriples* (`http://xtriples.spatialhumanities.de`).

XTriples makes it possible to extract RDF statements out of any HTTP based XML repository using a simple configuration based on statement patterns. The webservice's guiding principles are:

- Generic ⇒ works on any XML
- Simple ⇒ easy to configure
- Powerful ⇒ for building complex statements with proper ontology support
- Flexible ⇒ returns several formats
- RESTful ⇒ uses `http` for request and response
- Location-independent ⇒ freely accessible via the WWW
- Platform-independent ⇒ server-side processing
- Customisable ⇒ adaptable to a project's needs

The webservice is able to crawl *any* XML file, extract semantic statements and subsequently return them in a specified format (cf. fig. 1 on p. 58).

The webservice can be used with direct POST, form-style POST or GET requests. The required statement patterns are passed to the webservice in the form of a simple XML configuration based on XPATH/XQuery expressions. The triples are constructed based on the instructions given in the configuration (cf. listing 3, the documentation under `http://xtriples.spatialhumanities.de/documentation.html` as well as the exemplary data sets under `http://xtriples.spatialhumanities.de/examples.html`).

During the extraction it is possible to reach beyond the boundaries of a specific XML repository and include external XML resources or data units from these resources on-the-fly. This way norm data from the *German National library*'s authority file (GND), data from *DBpedia*, or any other third party can be included. Furthermore, the configuration can be used to fine tune the quantity, the type, and the granularity of statements the webservice is supposed to extract.

Listing 3: Sample structure of an XTriples configuration

```
<xtriples>
    <configuration>
        <vocabularies>
            <vocabulary prefix="tei" uri="http://www.tei-c.org/ns/
                1.0"/>
            <vocabulary prefix="cs" uri="http://www.bbaw.de/
                telota/correspSearch"/>
            [...]
            <vocabulary prefix="foaf" uri="http://xmlns.com/foaf/
                0.1/"/>
            [...]
        </vocabularies>
        <triples>
            <statement>
                <subject>//tei:correspAction[@type="sent"]/
                    tei:persName/@ref</subject>
                <predicate prefix="rdf">type</predicate>
                <object prefix="foaf" type="uri">Person</object>
            </statement>
             [...]
            <statement>
                <subject>//tei:correspAction[@type="sent"]/
                    tei:persName/@ref</subject>
                <predicate prefix="rdfs">label</predicate>
                <object type="literal" lang="de">//tei:correspAction
                    [@type="sent"]/tei:persName/text()</object>
             </statement>
        </triples>
    </configuration>
        <collection uri="http://correspSearch.bbaw.de/api/v1/
            tei-xml.xql?correspondent=http://d-nb.info/gnd/
            118540238&startdate=1793-01-01&enddate=1808-02-02">
            <resource uri="{//tei:correspDesc}"/>
        </collection>
</xtriples>
```

(Below `<collection>` the XML data that should be processed by the service is configured. Below `<vocabularies>` it is possible to configure the RDF vocabularies that should be used. Below `<triples>` the `<statement>` patterns that contain the *subjects*, *predicates* and *objects* are configured. The notation will output the following in `turtle` syntax: `cs:personID rdf:type foaf:Person ; rdfs:label "Johann Wolfgang von Goethe"@en .`)

The implementation as a webservice has the clear advantage that the user does not need to install any software besides a web browser in order to translate

data from XML to RDF. XTriples can also be used as an (external) RDF interface (like a proxy) for one or more XML based repositories. The prerequisite for this kind of usage is simply a repository's availability via HTTP.

The webservice can return the following formats: RDF/XML, Turtle, NTriples, NQuards, Trix, JSON, SVG, and custom XML for debugging purposes.

The result of an *XTriples* extraction is thus available in a wide variety of RDF-serialisations (cf. fig. 2 on p. 59). As seen above, besides the purely RDF based format, a repository's semantic relations can also be written to an SVG file or piped to other Semantic Web tools.[2]

XTriples was developed in the context of the long term research project *Deutsche Inschriften* (German Inscriptions, currently carried out by six of the eight German Academies of Sciences) together with the project *Inschriften im Bezugssystem des Raumes* (Inscriptions in their Spatial Context, funded between 2012–2015 by the German Ministry of Education and Research). The software is released on Github (`https://github.com/spatialhumanities/xtriples`) in a stable version under MIT license and comes with a detailed technical documentation (`http://xtriples.spatialhumanities.de/documentation.html`).

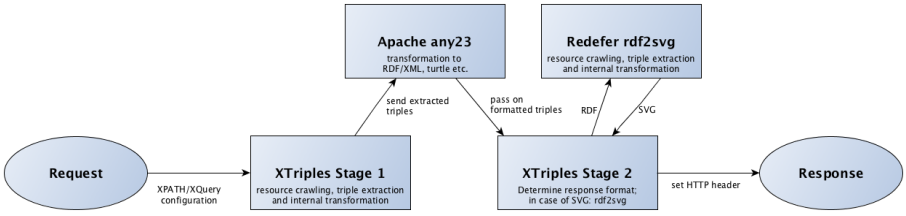

Fig. 1. Flow chart showing *XTriple*'s processing architecture.

3 Exemplary Use Cases

The *XTriples* webservice is currently used in different projects associated with or located at the *Academy of Sciences and Literature* in Mainz. Section 3.1 will focus on the the extraction of RDF statements from XML in a use case from the field of spatial humanities. Section 3.2 will exemplify the webservice's capabilities within the contexts of three further digital humanities projects and thus illustrate the webservice's flexibility in different project environments and its adaptiveness to their varying research questions.

[2] It is for example possible to pass the data to the *RDF-to-SVG* webservice (`http://www.rhizomik.net/html/redefer/rdf2svg-form`) or to the RDF visualisation library *d3sparql* (`http://biohackathon.org/d3sparql`). The *XTriples* URL can also be passed as parameter to *Visual RDF*, e.g `https://graves.cl/visualRDF/?url=xtriples.spatialhumanities.de/extract.xql?configuration=YourXTripleConfig.xml`.

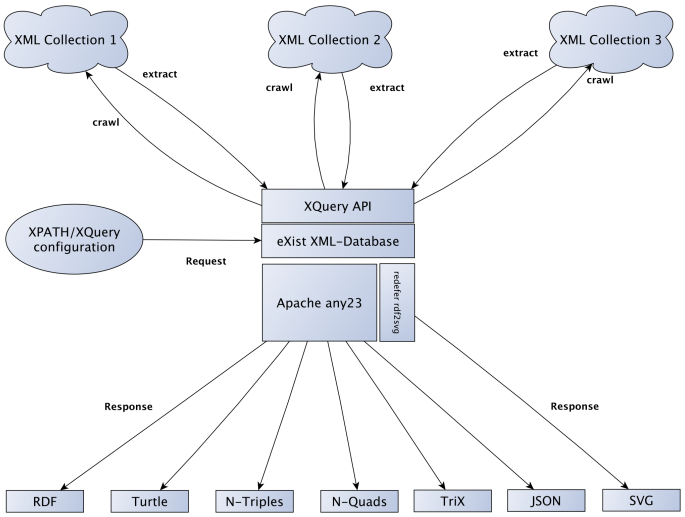

Fig. 2. Diagram showing *XTriple*'s modular structure.

3.1 The Projects "German Inscriptions Online" and "Inscriptions in their Spatial Context"

The long term research project *Deutsche Inschriften* (DI) is a joint undertaking of six German Academies of Sciences and the Austrian Academy of Sciences. The research focuses on collecting, editing, and interpreting medieval and early modern Latin and German inscriptions. They often occur in conjunction with figurative elements or spatial as well as architectural features. The inscriptions themselves are mostly in medieval Latin or in historical or regional varieties of the German language. The geographical area of research consists of Germany, Austria, and South Tyrol. The inscription records range from approximately 500 AD to 1650 AD [1, 6, 12]. The project's scholars carry out their research within a wide scope of interests ranging from art history, philology, and linguistics to the history of ideas. The research results are published in 90 volumes. More than 43 of these volumes, including over 17.000 records, are currently accessible through the online database *Deutsche Inschriften Online* (DIO) (German Inscriptions Online, `http://www.inschriften.net/`).

The *Federal Ministry of Education and Research* (BMBF) funded project *Inscriptions in their Spatial Context* (IBR, 2012–2015) had the aim to combine and analyse spatial data from terrestrial laser scanning with the epigraphical data made available by DIO. To achieve this, relevant parts of the epigraphical data (marked up in TEI-EpiDoc) had to be made available as RDF. This was done with the *XTriples* webservice. To achieve this, IBR first compiled its own semantic predicates in a project specific ontology (see [5] for further information about the underlying conceptual design).

The IBR project then harvested relevant TEI-EpiDoc records from DIO's epigraphical database via its REST interface according to guidelines formerly worked out by the project's participants. Afterwards the harvested EpiDoc corpus was processed by the *XTriples* webservice and translated to RDF. The transformation was very efficient regarding time and labour as well as highly suited to the project's needs. It was possible to incorporate the project's custom-made ontology. This high level of adaptability was due to *XTriples* inherent customisability via its flexible configuration.[3] Subsequently the RDF data was loaded into the project's main working environment, the *Generic Viewer* [8]. Within the viewer the researchers were able to further analyse the semantic connection between epigraphical objects and their spatial environment.

Using this synthesized data approach, IBR was able to allocate spatial segments of a church's interior to distinct socio-political groups. Besides this, spatial areas within the church could be identified and annotated with their respecitive liturgical, ritual, or social function, thereby highlighting "potentially relevant factors like the social division of the congregation room, procession routes, and other places of liturgical practices" [7]. Additional textual or pictorial sources were linked to the already edited epigraphical material.

IBR's researchers were able to estimate the former installation or *in situ* location of epigraphical artefacts—for example an epitaph and retable endowed by the canon Petrus Lutern—by means of a *Viewshed*-analysis [8]. The IBR case study "illustrates that the line between visualisation and analysis in the humanities' qualitative research is a blurred one" [9].

This use case highlights the applicability of semantic annotations in regard to historical research within a highly technical environment. But the concept could also be very useful in regard to more public-oriented spatial visualisations [7, 9], like the virtual tour through the nave of St. Michael's church in Hildesheim [11]. The textual data contained within such virtual tours of world heritage sites could be processed in a first step and in a subsequent second step linked to items in repositories that contain semantically enriched historical data—like e.g. *Europeana* [8]. This would provide the means for further study and computational analysis of data sets from the humanities.

3.2 Other Projects

Besides the projects DI and IBR, *XTriples* is currently used by several other academic projects, namely the *Regesta Imperii*[4] and the *Schule von Salamanca*.[5] The *Salomon Ludwig Steinheim Institut für deutsch-jüdische Geschichte*[6] uses *XTriples* for a CIDOC-CRM based modelling of its EpiDoc-data. Together with

[3] At this stage it would also have been possible to process the data further, e.g. by visualising the semantics with SVG or by further enriching the RDF file with incorporated external sets.

[4] `http://www.regesta-imperii.de`

[5] `http://www.salamanca.school/`

[6] `http://www.steinheim-institut.de`

the *Digital Humanities* departement of the *Berlin-Brandenburg Academy of Sciences* (BBAW) a joint effort is under way to develop a common API between *XTriples* and *correspSearch*, the BBAW's decentralized aggregation-tool for metadata of digital scholarly editions of correspondence.

The following list gives a first insight into some of the diverse use cases of XTriples:

1. Extraction of semantic relations and subsequent visualisation of the familial ties present within the data made available through the *Steinheim Institute's* EpiDat corpus.[7] The data represents the Jewish cemetery in Hamburg-Altona.[8]
2. Extraction of semantic relations from Johann Wolfgang von Goethe's correspondence and subsequent SVG visualisation of a specific sub-network. The underlying data was retrieved from CMI records aggregated by *correspSearch*. Norm data provided by *Geonames* and the DNB's authority file was incorporated on-the-fly via their respective RDF interfaces.[9]
3. Exemplary extraction of semantic relations from *correspSearch's* CMI/TEI database and subsequent visualisation of European communication networks implicitly present within the records (fig. 3 on p. 62).[10]

More use cases demonstrating particular functionalities of *XTriples* can be found on the *XTriples* website. The following presentation gives a quick overview of the service's capabilities and modes of operation.[11]

4 Conclusion

As shown above Semantic Web tools that wish to step up to the challenges posed by humanities research have to address many different topics. When working with data sets from the humanities the huge diversity found in this field of research always has to be taken into account. Generally speaking it is hard to find even two digital scholarly editions that are structurally comparable or pose the same research questions, even though they may be working on the same topic. Any semantic web tool that should be usable for humanities research and data must therefore be able to deal with this surprisingly high degree of diversity. It should not be confined to a fixed set of ontologies, schemata, or output formats, but permit the highest possible degree of flexibility and thereby further "a spirit of openness shared by all parties involved" ([3] par. 32) in return.[12]

[7] `http://www.steinheim-institut.de/cgi-bin/epidat`

[8] `http://xtriples.spatialhumanities.de/examples/dh/epidat/index.html`

[9] `http://xtriples.spatialhumanities.de/extract.xql?configuration=http://xtriples.spatialhumanities.de/examples/dh/correspSearchLetters.xml&format=svg`

[10] `http://metacontext.github.io/presentation-correspsearch-xtriples/viz/map.html`

[11] `http://metacontext.github.io/presentation-correspsearch-xtriples`

[12] This degree of flexibility also has to include the use of open licences in order to grant legal safeguards for re-use and modification of program code and data.

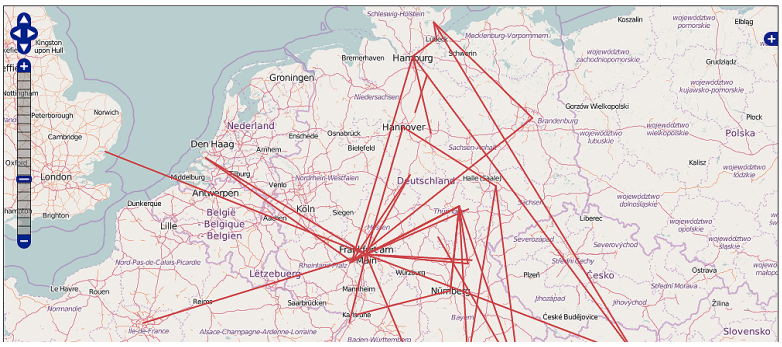

Fig. 3. Visualisation of the correspondence between Johann Wolfgang von Goethe, Carl Maria von Weber and Samuel Thomas von Soemmering.

Thinking in terms of monolithic software is not enough when it comes to bridging the gap between modelers with different approaches (see [2] par. 20, 43–44) as well as between the humanities and the Semantic Web. The experiences made during the projects discussed above clearly show that this approach will not be successful in the end. It might be much more realistic to apply a microservices based approach to this huge task, where each tool can focus on it's particular strength and provide an (ideally powerful) interface that other tools can use and customise for their own needs. It almost goes without saying that the software tools and data sets also have to be accessible under open conditions, ideally with free licenses allowing third parties to build upon the existing work.

To summarise, it is appropriate to say that in the same way as the Semantic Web distinguishes itself by simultaneously being distributed and heterogenous and, at the same time, connected and analysable, the tools and the researchers trying to utilise Semantic Web technologies for the humanities have to be highly flexible in their approaches, too. It may be worthwile to change the focus from an all-in-one to a more project and microservices oriented approach that scales much better to the semantic potentials concealed in the many different data repositories from the humanities that are available online.

References

1. Brandi, K.: Grundlegung einer deutschen Inschriftenkunde. In: Deutsches Archiv für Erforschung des Mittelalters Bd. 1 (1937), pp. 11–43. `http://www.digizeitschriften.de/dms/img/?PID=PPN345858700_0001%7Clog10` (retrieved 01/2016)
2. Eide, Ø.: Ontologies, data modelling, and TEI. In: Journal of the Text Encoding Initiative, Issue 8 (Dec. 2015–2015). `http://jtei.revues.org/1191` (retrieved 10/2015)
3. de la Iglesia, M., Göbel, M.: From entity description to semantic analysis: The case of Theodor Fontane's notebooks. In: Journal of the Text Encoding Initiative, Issue 8 (Dec. 2015–2015). `http://jtei.revues.org/1253` (retrieved 10/2015)

4. de la Iglesia, M., Moretton N., Brodhun, M.: Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten. In: Neuroth, H., Rapp, A., Söring, S. (eds.) TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. Göttingen (2015), pp. 91–102. DOI: http://dx.doi.org/10.3249/webdoc-3947 (retrieved 10/2015)
5. Haft, M.: RDF als Verknüpfungsmethode zwischen geisteswissenschaftlichen Forschungsdaten und Geometrien am Beispiel des Projektes "Inschriften im Bezugssystem des Raumes". In: Skriptum 2, Nr. 3. http://nbn-resolving.de/urn:nbn:de:0289-2013120622 (retrieved 10/2015)
6. Kloos, R. M.: Die Deutschen Inschriften. Ein Bericht über das deutsche Inschriftenunternehmen. In: Studi medievali Ser. 3, Vol. 14 (1973), pp. 335–362
7. Lange, F., Unold, M.: Relating Texts to 3D-Information: A Generic Software Environment for Spatial Humanities. Lausanne (2014). http://dharchive.org/paper/DH2014/Paper-680.xml (retrieved 01/2016)
8. Lange, F., Unold, M.: Semantisch angereicherte 3D-Messdaten von Kirchenräumen als Quellen fr die geschichtswissenschaftliche Forschung. In: Baum, C., Stäcker, T. (eds.) Grenzen und Möglichkeiten der Digital Humanities (= Sonderband der Zeitschrift fr digitale Geisteswissenschaften, 1) Wolfenbüttel (2015), text/html Format. DOI: 10.17175/sb001_015
9. Lange, F., Schwartz, F., Unold, M.: GenericViewer – Semantische Annotation und 3D-Informationen in den Spatial Humanities. Passau (2014). https://i3mainz.hs-mainz.de/sites/default/files/public/data/Lange-GenericViewer_-_Semantische_Annotation_und_3D-Informationen-2421034.pdf (retrieved 01/2016)
10. Lanthaler, M.: Third Generation Web APIs. Bridging the Gap between REST and Linked Data. Diss. Institute of Information Systems and Computer Media. Technische Universität Graz (2014). http://www.markus-lanthaler.com/research/third-generation-web-apis-bridging-the-gap-between-rest-and-linked-data.pdf (retrieved 10/2015)
11. Neovesky, A., Peinelt, J.: A Virtual Tour to the Inscriptions of the UNESCO World Heritage Site St. Michael in Hildesheim. In: Electronic Visualisation and the Arts (EVA 2015) Conference Proceedings. London (2015), DOI: 10.14236/ewic/eva2015.31
12. Nikitsch, E. J.: Fritz V. Arens als Mainzer Inschriftensammler und Epigraphiker. In: Mainzer Zeitschrift Vol. 103 (2008), pp. 231–243.
13. Polleres, A., at. al.: XSPARQL Language Specification. Galway (2009). http://www.w3.org/Submission/xsparql-language-specification (retrieved 10/2015)
14. Schrade, T.: Datenstrukturierung. In: Frietsch, U, Rogge, J. (eds.) Über die Praxis des kulturwissenschaftlichen Arbeitens. Ein Handwörterbuch. Bielefeld (2013), pp. 91–97